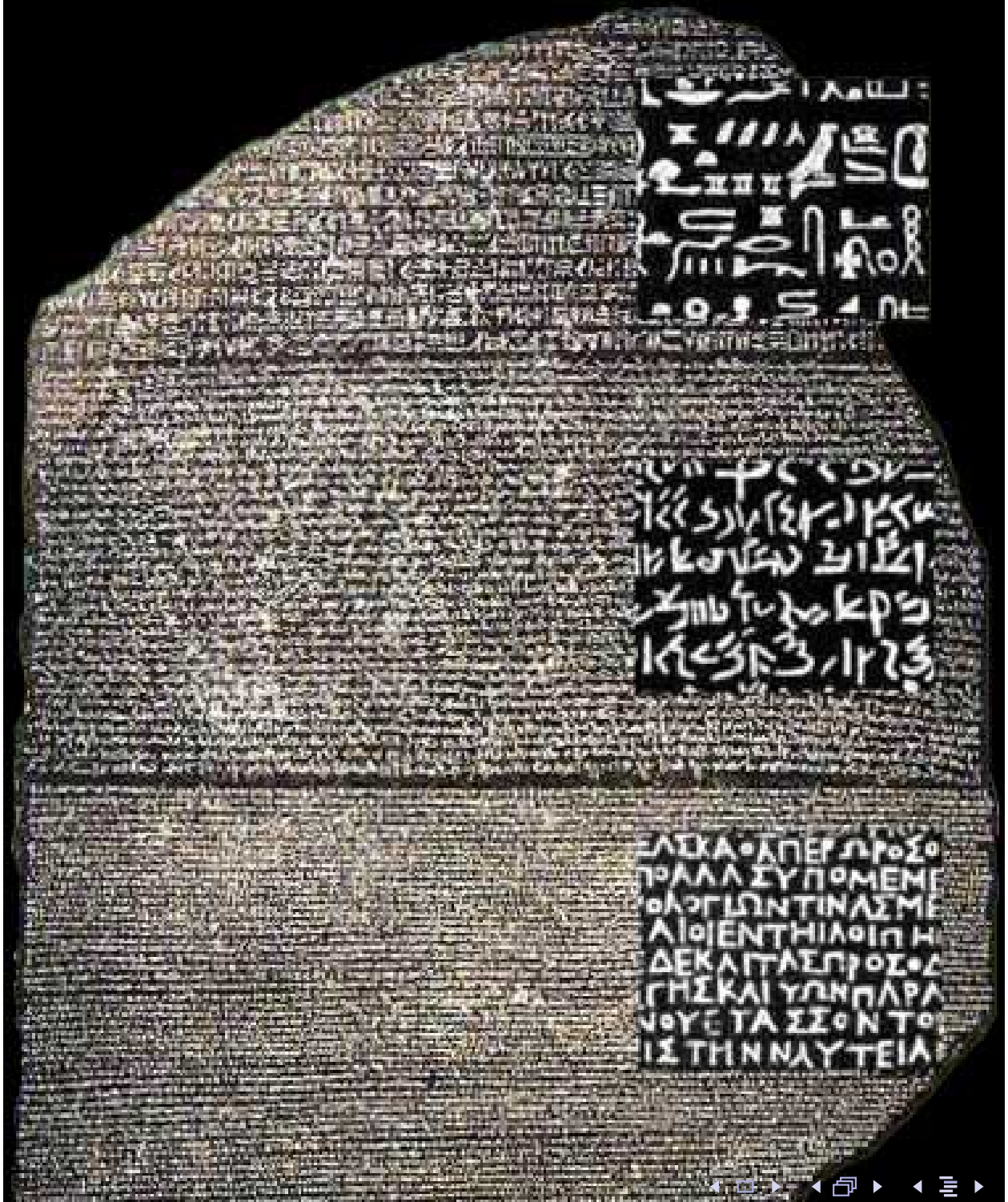


/ Domů / Výběr korpusů / Dotaz Michal Křen | Odhlášení | nápověda |

1-6 z 6 (stránka 1 z 1 >>)

intercorp_cs *(74068 tokenů)*	intercorp_bg *(74274 tokenů)*	intercorp_en *(80252 tokenů)*	intercorp_pl *(67660 tokenů)*
Zobraz možnosti Filtr Kwic	Kwic		
Rowling, J.K.,Harry Potter a Kámen mudrců Teta Petunie se napřed Harryho rozzlobeně zeptala , jestli toho člověka zná , a pak ho i s Dudleym spěšně odtáhla z krámu , aniž něco koupila .	Rowlingová, J,Хари Потър и философският камък След като попита яростно Хари дали познава този мъж , леля Петуния ги поведе бързо навън , без да купи нищо .	zobraz kontext Rowling, J.K.,Harry Potter and the Sorcerer's Stone After asking Harry furiously if he knew the man , Aunt Petunia had rushed them out of the shop without buying anything .	zobraz kontext Rowling, Joanne K.,Harry Potter i kamień filozoficzny Ciotka Petunia zapytała Harry'ego ze złością , czy zna tego człowieka , a potem wygoniła ich ze sklepu , choć niczego jeszcze nie kupiła .
Rowling, J.K.,Harry Potter a Kámen mudrců Hermiona vystřelila ruku tak vysoko , jak jen bylo možné , aniž by přitom vstala ze sedačky , Harry však neměl sebemenší tušení , co to bezoár je .	Rowlingová, J,Хари Потър и философският камък Хърмаяни протегна ръка толкова високо , колкото можеше , без да стане от стола си , но Хари нямаше и най-бледа представа какво е безоар .	zobraz kontext Rowling, J.K.,Harry Potter and the Sorcerer's Stone Hermione stretched her hand as high into the air as it would go without her leaving her seat , but Harry did n't have the faintest idea what a bezoar was .	zobraz kontext Rowling, Joanne K.,Harry Potter i kamień filozoficzny Hermiona wyciągnęła rękę tak wysoko , jak zdołała , nie opuszczając swojej ławki , ale Harry nie miał zielonego pojęcia , co to jest bezoar .
Rowling, J.K.,Harry Potter a Kámen mudrců Nasedl na koště , odrazil se , jak mohl nejvíc , a pak už se řítil vzhůru , vítr mu svištěl ve vlasech a jeho hábit vlál za ním - potom si v návalu divoké radosti uvědomil , že objevil něco , co umí , aniž by ho to někdo musel učit - bylo to snadné , bylo to úžasné ! Trochu koště nadzdvihl , aby se dostal ještě výš , a zdola slyšel vřeštění a jíkání děvčat a Ronovo obdivné zavýsknutí .	Rowlingová, J,Хари Потър и философският камък Възседна метлата си , отблъсна се силно от земята и се понесе нагоре . Въздухът свистеше в косата му , одеждите му плющяха зад него и в прилив на бясна радост той осъзна , че е открил нещо , което можеше да прави , без да го учат - това беше лесно , това беше прекрасно .	zobraz kontext Rowling, J.K.,Harry Potter and the Sorcerer's Stone He mounted the broom and kicked hard against the ground and up , up he soared ; air rushed through his hair , and his robes whipped out behind him - and in a rush of fierce joy he realized he 'd found something he could do without being taught -- this was easy , this was wonderful . He pulled his broomstick up a little to take it even higher , and heard screams and gasps of girls back on the ground and an admiring whoop from Ron .	zobraz kontext Rowling, Joanne K.,Harry Potter i kamień filozoficzny Dosiadł miotły i odepchnął się mocno od ziemi . Wystrzelił w powietrze jak pocisk ; pęd rozwiał mu włosy , a szata łopotała za nim jak żagiel . Poczuł gwałtowny przypływ dzikiej radości - uświadomił sobie , że robi coś , czego nigdy się nie uczył , że to bardzo łatwe i ... cudowne . Zadarł lekko kij , żeby wznieść się jeszcze wyżej , a z ziemi usłyszał podniecone wrzaski dziewczyn i donośny okrzyk podziwu Rona .
Rowling, J.K.,Harry Potter a Kámen mudrců Zhltal večeři , aniž by si všiml , co vlastně jí , a pak se spolu s Ronem hnali nahoru , aby Nimbus Dva tisíce konečně vybalili .	Rowlingová, J,Хари Потър и философският камък Той излапа вечерята си , без да забелязва какво яде , и след това се втурна с Рон нагоре , за да разопакова най-после своята « Нимбус две хиляди » .	zobraz kontext Rowling, J.K.,Harry Potter and the Sorcerer's Stone He bolted his dinner that evening without noticing what he was eating , and then rushed upstairs with Ron to unwrap the Nimbus Two Thousand at last .	zobraz kontext Rowling, Joanne K.,Harry Potter i kamień filozoficzny Po lekcjach wchłonął szybko kolację , nie bardzo wiedząc , co połyka , i popędził z Ronem na górę , żeby rozwinąć Nimbusa Dwa Tysiące .
Rowling, J.K.,Harry Potter a Kámen mudrců	Rowlingová, J,Хари Потър и философският камък	zobraz kontext Rowling, J.K.,Harry Potter and the Sorcerer's Stone	zobraz kontext Rowling, Joanne K.,Harry Potter i kamień filozoficzny

Export: xls Pohled: horizontální

Sketch Engine beta

About Home Settings Change password Log out

Search in Help

user: Mr. Alexandr Rosen corpus: EUROPARL5, English-German Search in EUROPARL5, English-Ger

Concordance
Word List
Word Sketch
Thesaurus
Find X
Sketch-Diff

Save
View options
KWIC/Sentence
Sort
Left | Right
Node
References
Shuffle
Sample
Filter
Frequency
Node tags
Node forms
Doc IDs
Collocations
ConcDesc

Corpus: **EUROPARL5, English-German**
Hits: **40** (0.9 per million)

Page 1 of 2 Go Next | Last

	EUROPARL5, English-German	EUROPARL5, German-English
doc#116	They have no owners and should be caught and sterilised <g/> , but <g/> , instead <g/> , they are the victims of indiscriminate hunting by the people of the city <g/> , who receive a reward for every dog they kill and whose body they take to the mayor as proof <g/> .	Wir sprechen hier über Tiere, und es ist mir ein dringendes Bedürfnis, in diesem Zusammenhang zu erwähnen, dass die ausgesetzten und herrenlosen Hunde in Bukarest, die eigentlich eingefangen und sterilisiert werden müssten, rücksichtslos von den Bürgern gejagt werden, die eine Prämie für jeden getöteten Hund bekommen, dessen Kadaver sie dem Bürgermeister bringen.
doc#188	Stop letting the tail wag the dog <g/> , and lead the Netherlands back to the heart of European decision-making <g/> , as befits the founder status of your country in our Union <g/> .	Hören Sie auf, den Schwanz mit dem Hund wedeln zu lassen und führen Sie die Niederlande wieder ins Herz der europäischen Entscheidungsprozesse, wie es sich für den Status Ihres Landes als Gründungsmitglied unserer Union geziemt.
doc#623	Meanwhile they decided that it had to be ratified unanimously <g/> , exactly like the gardener <g/> 's dog that neither eats cabbage itself nor lets anybody else <g/> .	Inzwischen wurde auch noch beschlossen, dass er einstimmig ratifiziert werden muss, wie bei des Gärtners Hund , der keinen Kohl frisst, aber auch nicht will, dass andere davon essen.
doc#708	I know that here in Europe and in certain continents <g/> , man <g/> 's best friend <g/> , the dog <g/> , has a passport in order to be able to travel from one country to another <g/> .	Ich weiß, dass hier in Europa und auf einigen Kontinenten der beste Freund des Menschen, der Hund , einen Pass besitzt, um von einem Land in ein anderes reisen zu können.
doc#766	I have a dog myself <g/> , and he has told me that he does not want a microchip or a tattoo <g/> .	Ich habe auch einen Hund , und der hat mir gesagt, er will weder einen elektronischen Mikrochip noch tätowiert werden.
	You see these poor dogs on aeroplanes <g/> ,	Wenn man bei Flugreisen sieht, wie diese armen Hunde bellend, wimmernd, winselnd in diese grässlichen Käfige

Hits: **74** | i.p.m. : 0.35 (related to the whole InterCorp v7 - Czech) | ARF : **3.40** | ! | Result is shuffled 1 / 2

	InterCorp v7 - Czech	InterCorp v7 - English	InterCorp v7 - Polish
☐	Právě když byla večeře hotova , někdo zazvonil , Švejk **šel otevřít** , vrátil se za chvíli a hlásil :	Just when dinner was ready , the door bell rang . Švejk went to the door , then came back and reported : " He 's here again , Field Chaplain , sir .	W chwili gdy kolacja była gotowa , ktoś zadzwonił . Szwejk pośpieszył otworzyć drzwi i po chwili zameldował :
☐	" Tak abychom už **šli hledat** ten polní oltář , " vybízel Švejk , " je už ráno . "	" Then , we should be looking for the field altar now , " Švejk challenged . " It 's already morning . "	- Może by śmy wreszcie poszli na poszukiwanie ontarza polowego - zapraszał Szwejk - już jest ranek .
☐	Pak se **šel umýt** do koupelny .	Then he went to wash in the bathroom .	Potem poszedł umyć się do łazienki .
☐	A hned **šel žádat** velitele o dva dny volna , jenomže velitel se mu je zdráhal dát , protože právě v té době byly na Stáňu z dolů a z kasáren samé stížnosti , které si způsoboval roztěkaností a podrážděností .	He tried to get a two-day pass from the company commander , but found him none too willing : he had been getting nothing but complaints about Stana from both the mines and the barracks , complaints about his absentmindedness and irritability .	I natychmiast poszedł do komendanta , żeby mu dał dwa dni urlopu , tylko że komendant wzdragał go się puścić , ponieważ właśnie w tym czasie przychodziły na Staszka z kopalni i z koszar same skargi , spowodowane jego roztargnieniem i zdenerwowaniem .
☐	Ženy z okolí **šly doprovázet** svoje muže na vojnu , a on věděl , že ti mužové určitě svým ženám slibují , že se nedají zabít pro císaře pána .	Women from the vicinity would see their men off to military service , and he knew that those men were certainly promising their wives that they would not let themselves be killed for the Lord Emperor .	Pan rotmistrz widywał , jak żony odprowadzały mężów wezwanych do wojska , i z góry już wiedział , że ci mężowie obiecywali żonom jak najuroczyściej , iż nie dadzą się zabić dla najjaśniejszego pana .
☐	Rozhovor byl nekonečný , Agnes a Paul opakovali stejné věty , ujišťovali Lauru svou láskou , prosili ji , aby zůstala s nimi , aby je neopouštěla , až jim nakonec slíbila , že vrátí revolver do zásuvky a **půjde spát** .	The conversation was endless , Agnes and Paul kept repeating the same sentences , they assured Laura of their love , they begged her to stay with them , not to leave them , until she finally promised to return the gun to the drawer and go to sleep .	Rozmowa ciągnęła się w nieskończoność ; Agnes i Paul powtarzali w kółko to samo , zapewniali Laurę o swej miłości , błagali , by z nimi pozostała i nigdy więcej ich nie porzucała , toteż przyrzekła wreszcie , że schowa rewolwer do szuflady i pójdzie spać .
☐	Telefonní rozhovory na vojně , to není žádné tlachání po telefonu , když někoho zveme , aby nás **šel navštívit** k obědu .	Telephone conversations in the army , they are no small talk over the phone , like when we ' re inviting somebody over to visit us for lunch .	Wojskowe rozmowy telefoniczne to nie pogawędka okazyjna , jak na przykład , gdy się kogoś przez telefon zaprasza na obiad .
☐	A jestli před devíti měsíci se **šla podívat** bez vás do varieté na atletické zápasy , kde vystupoval nějakej černoch , tu myslím , že by vám to přeci jen trochu vrtalo hlavou . "	And if nine months ago she went without you to a variety show to watch an athletic competition where some black guy performed , here I must think that there would still be a bug drilling through your mind a little after all . "	Jeśli przypadkiem przed dziewięcioma miesiącami zdarzyło się jej być w jakim " Variété " , gdzie odbywały się walki zapaśnicze , w których brał udział na przykład jakiś Murzyn , to przypuszczam , że w umyśle pańskim zakiełkowała by nieufność .
☐	Vyslovil jsem to - a už to **nešlo vzít** zpátky .	Once I 'd said it - there was no unsaying it .	Powiedział em na głos i już nie dało się tego cofnąć .
☐	" My jsme dnes , " zabreptal , " dostali rozkaz **jít fasovat** na cestu koňak .	" We have today , " he mumbled , " received an order to go and get cognac issued for the journey .	- Dostali śmy dzisiaj rozkaz - mówił szybko - żeby śmy poszli fasować koniak na drogę .
☐	" Ještě něco , Švejku , " otázal se nadporučík , když Švejk odcházel na poštu , " co je s tím psem , kterého jste **šel hledat** ? "	" One more thing , Švejk , " the Lieutenant said as Švejk was departing for the Post Office . " What 's with the dog that you went to look for ? "	- Jeszcze jedno , mój Szwejku - zawołał porucznik , gdy Szwejk wybierał się na pocztę . - Czy znaleźli ście jakiegoś psa dla mnie ?

The merits of a parallel corpus and how to get the most of it

Alexandr Rosen

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

Slavicorp 2018
Institute of the Czech National Corpus
24–26 September 2018

Why am I here?

- Maybe you'll see that a parallel corpus is what you really need.
- Maybe you'll tell me what you really need from a parallel corpus.

Outline

Outline

- Same text in multiple versions (languages, translations, ...)
- Alignment by text units
- Parallel texts are all around! They are useful for:
  - machine translation, information retrieval, projecting annotation, other NLP application
  - translators (CAT tools)
  - foreign language teaching
  - lexicographers
  - researchers
- Can a parallel corpus be useful to you?

Problems

- Authenticity
  - translationese
- Availability
  - not in all languages, genres, text types
  - legal restrictions
- Alignment
  - not 100%
- Specific tools
  - aligners
  - parallel concordancers

What a parallel corpus offers

- Translation is supposed to preserve meaning
- Parallel context
  - → explicit translation equivalence
  - → implicit annotation of meaning

- From meaning to form:
  - find equivalent forms in other languages or in the same language
  - – translation studies, contrastive linguistics, FLT, MT, CAT
- From form to meaning:
  - find meaning of a form through other languages
  - – text understanding, annotation projection, monolingual lexicography

What a parallel corpus needs ...

... in addition to alignment

- Annotation of linguistic meaning
  → linguistic analysis
- Explicit monolingual annotation, crosslingual comparison
  → implicit meaning equivalence
- A chance to make the implicit meaning equivalence explicit?
  → by shared annotation scheme and categories

Outline

Basics

- A part of the *Czech National Corpus*
- http://www.korpus.cz/intercorp/
- *2005 with much kudos to František Čermák
- At first as a service for linguistic and philological departments
- 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

History

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	

Content (release 11)

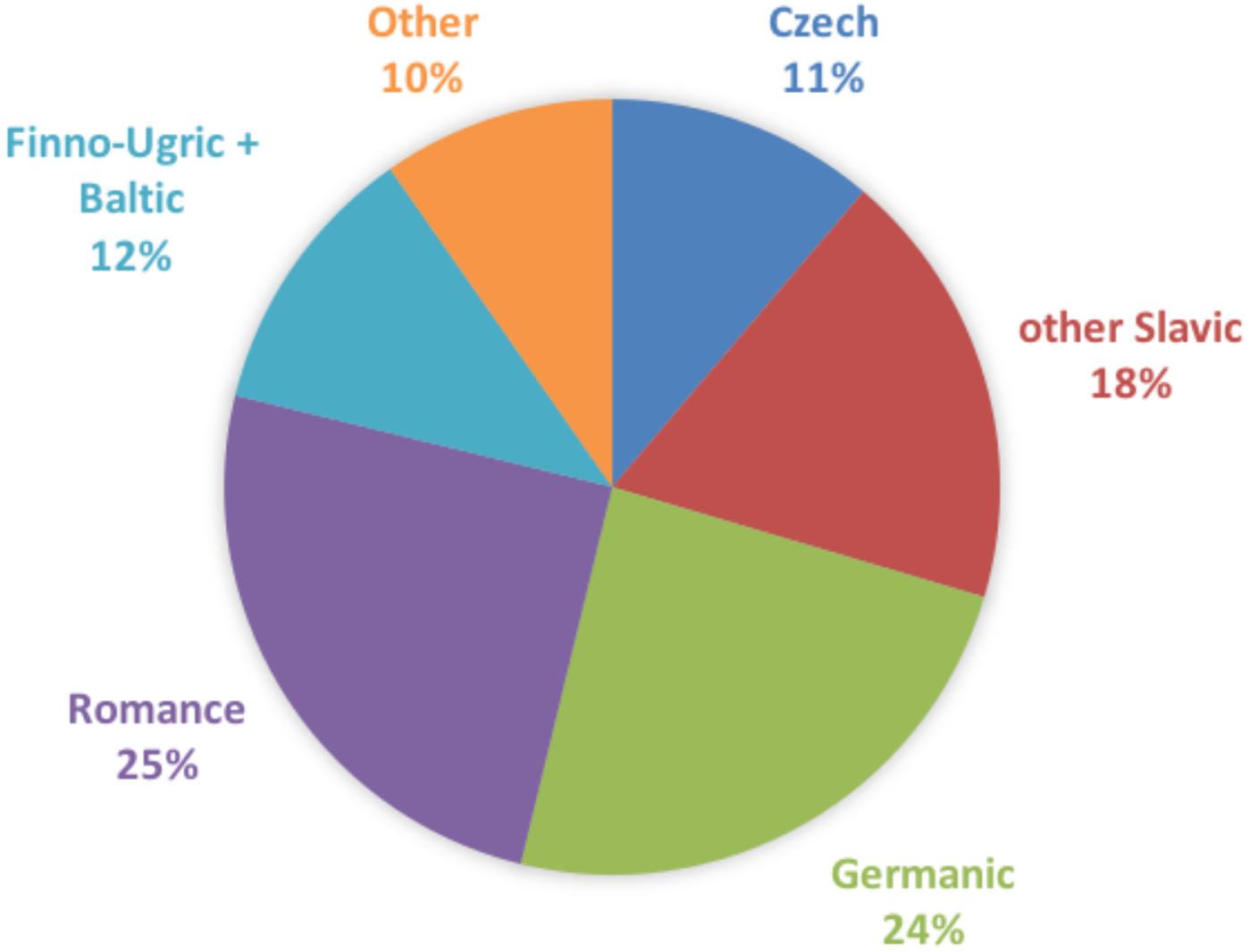
39 languages + Czech

- 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
- 11 other: ar, el, he, hi, ja, ms, mt, rn, sq, tr, vi

☛ Only few texts are available in more than 20 languages

☛ Languages differ wildly in the volumes of text

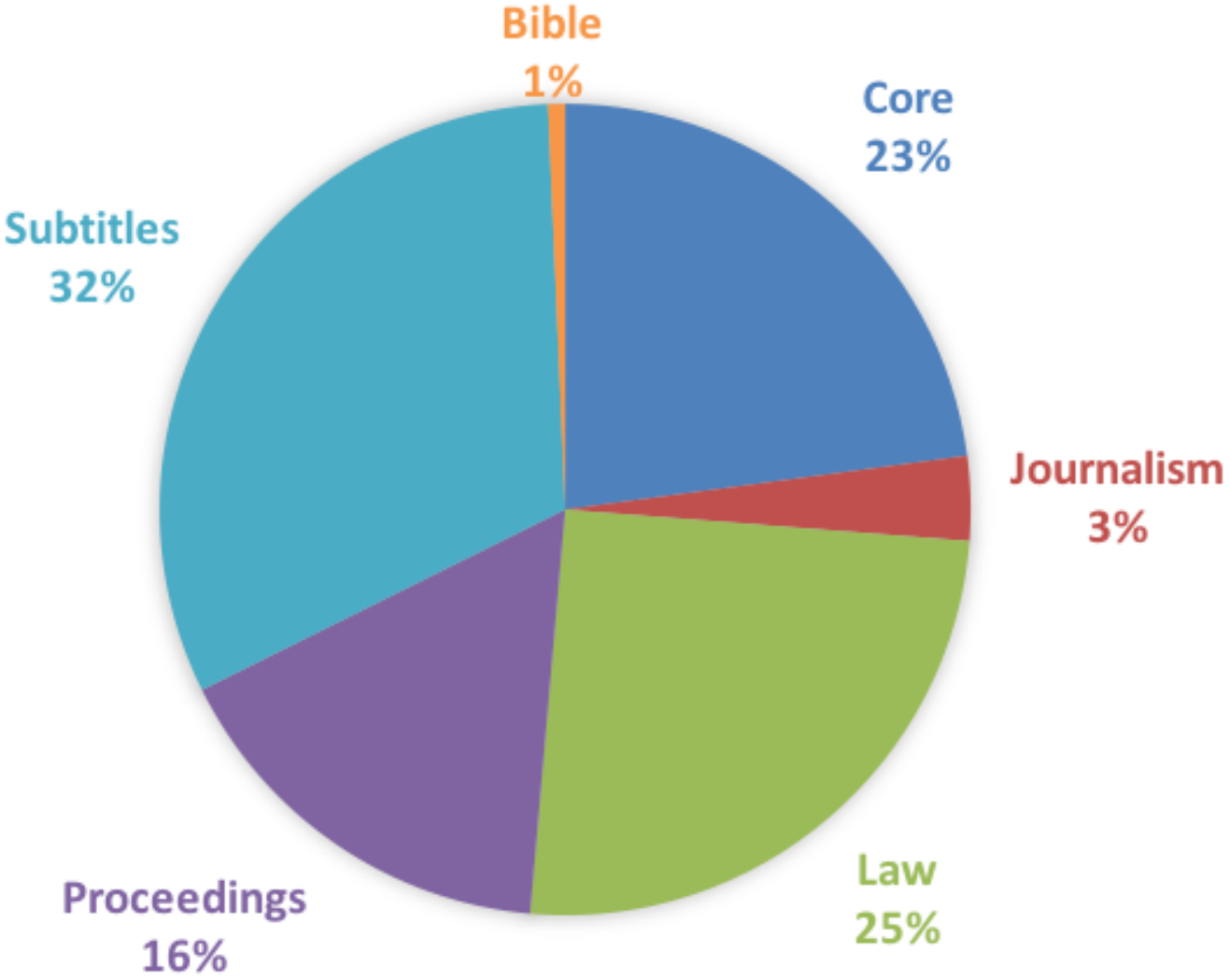
Language groups



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**

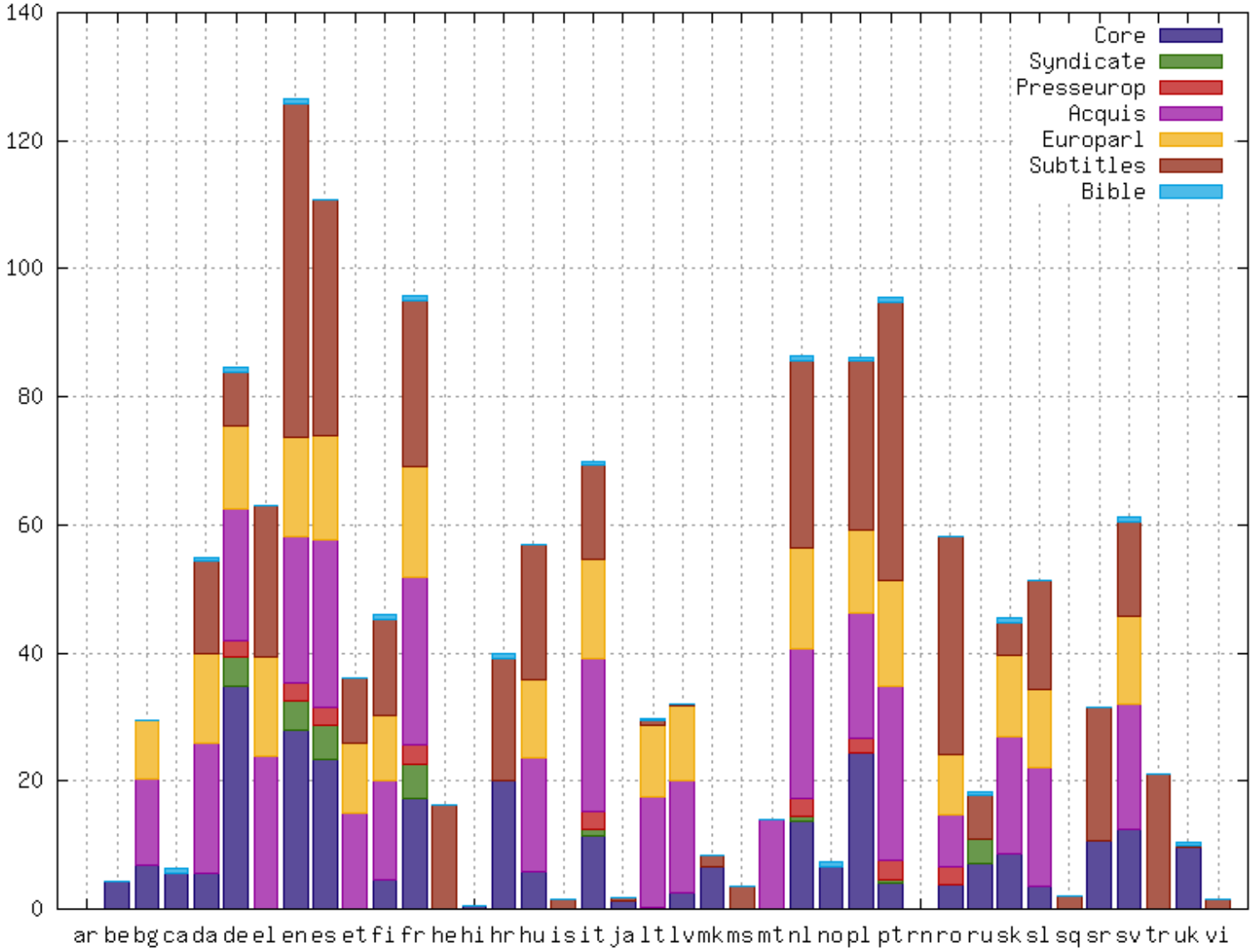
Text types


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

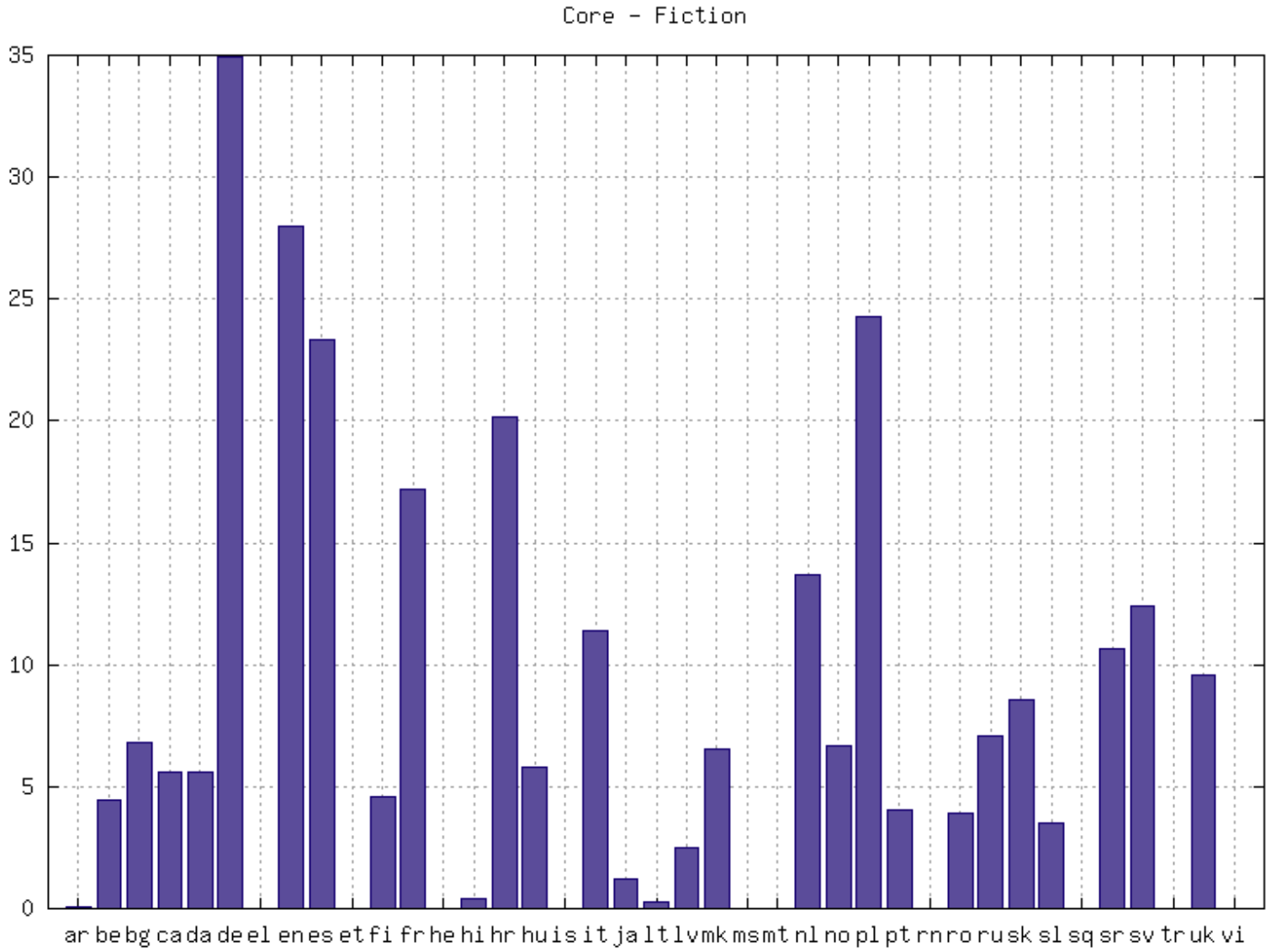
Size in million words

	Czech	Other Slavic	Other foreign	**Total**
Core	106.9	97.1	186.0	**390.0**
Journalism	6.4	6.1	44.6	**57.2**
Law	19.0	70.1	339.3	**428.5**
Proceedings	12.2	46.9	218.4	**277.5**
Subtitles	50.6	97.5	391.4	**539.5**
Bible	0.6	2.9	8.1	**11.6**
Total	195.8	320.6	1187.9	**1704.2**
No. of core texts	1,564	1,376	2,118	**5,058**

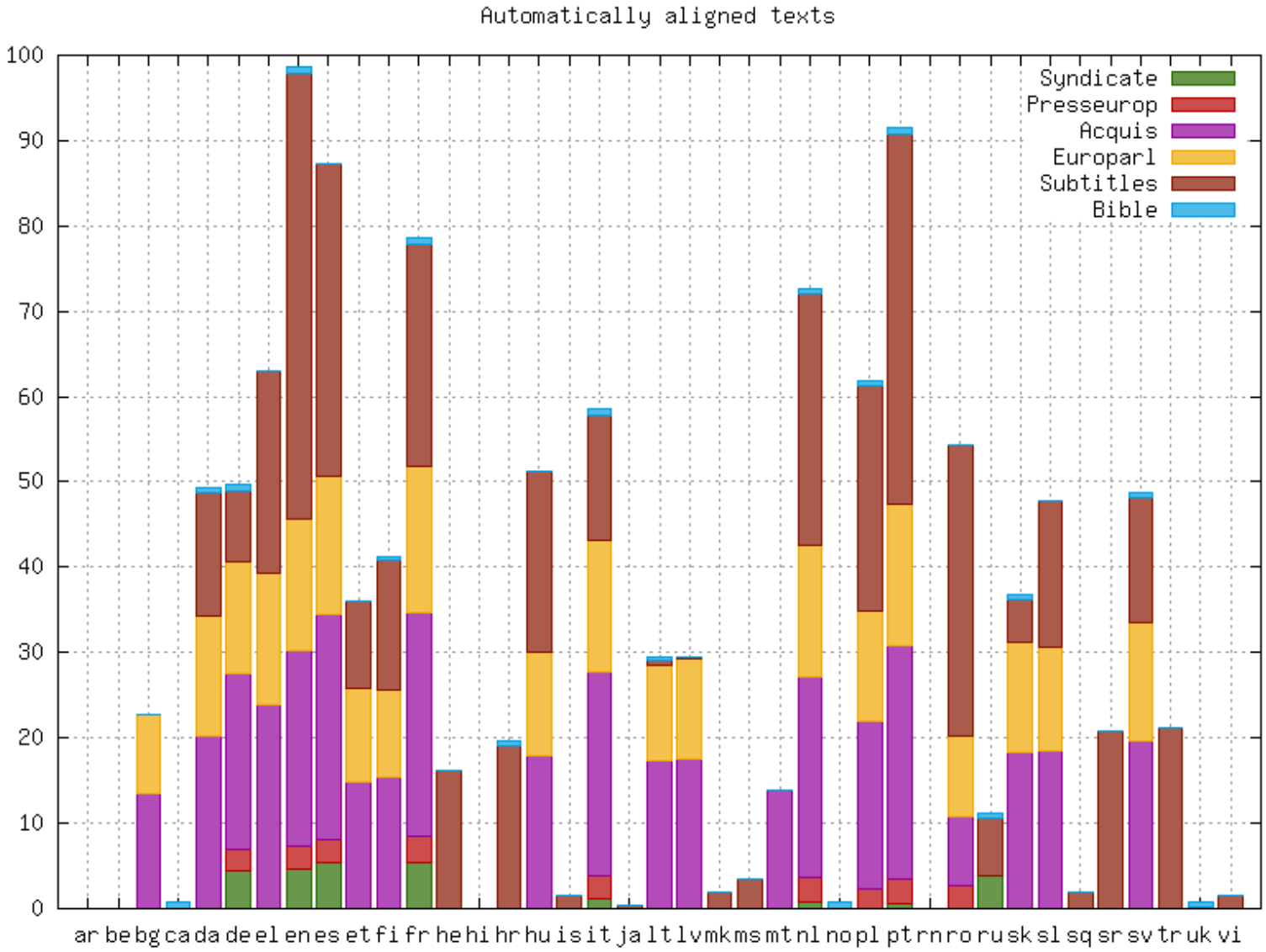
InterCorp by languages and text types



Core (mostly fiction)



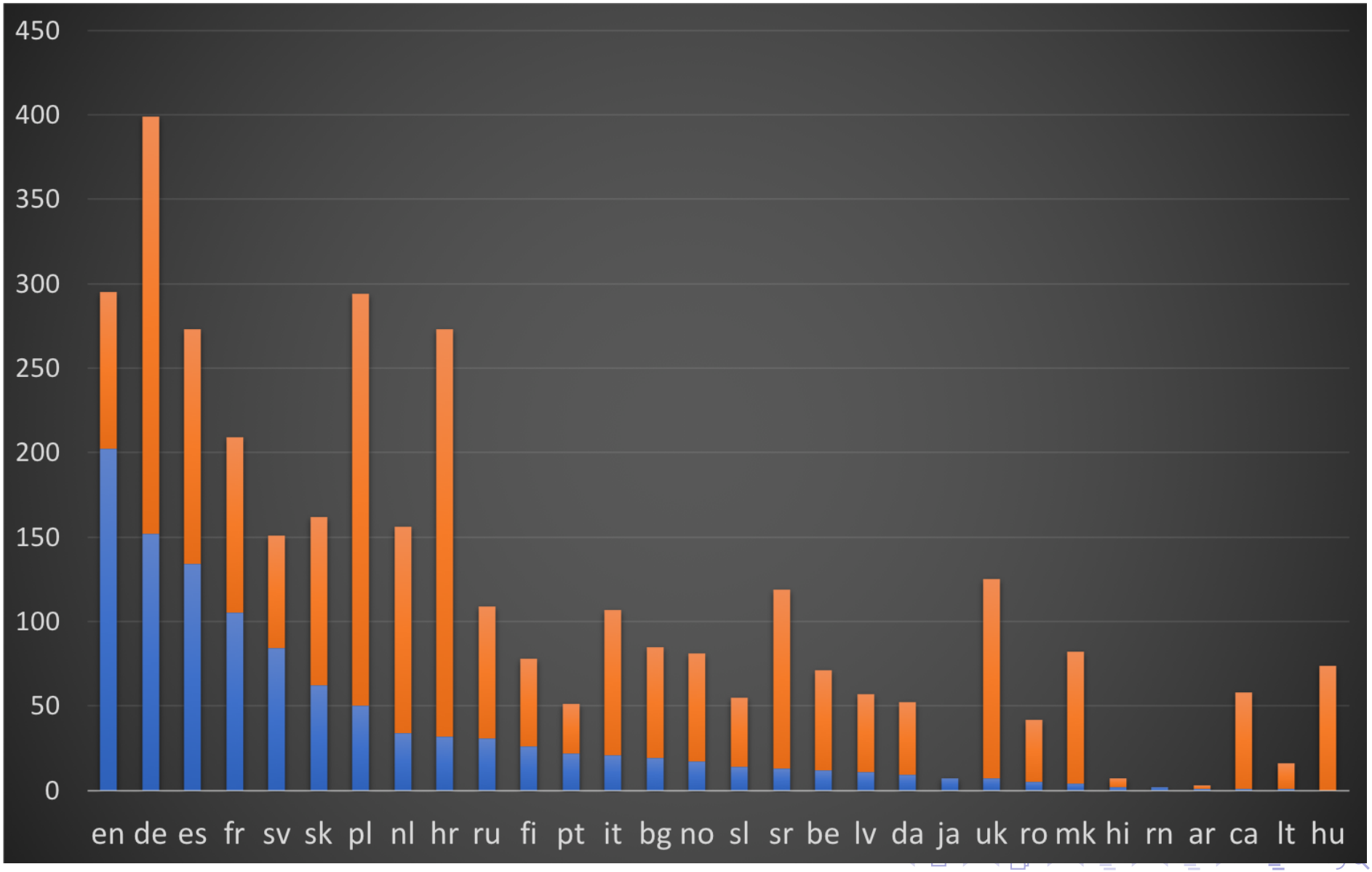
Collections (journalism, law, parliament proceedings)



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*
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...

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 CNC corpora
- Users take an active part

Outline

Concordances

KonText

http://kontext.korpus.cz

- Choice of *InterCorp* release
- Search filters:
  - languages, texts, publication year, text type
  - original/translation, language of the original
  - sex of the author and the translator
- Parallel queries, CQL
- Positive and negative filters on the concordances
- Export of concordances
- Sorting, frequency distribution, collocations
- Custom subcorpora

Lexical lookup

Treq – a database of translation equivalents

http://kontext.korpus.cz

- Pairs of lexical equivalents extracted from word-aligned parallel texts
- cs/en ↔ any other language
- Filtering by text groups
- Queries by forms or lemmas
- Support of regular expressions
- Single forms or multi-word expressions

Polish equivalents for Czech *bouře*

295	*burza*	≈ in printed dictionary
35	*śnieżyca*	
29	*sztorm*	≈ kinds of a storm: snow storm, on the sea
7	*zamieć*	
6	*zawierucha*	
13	*wichura*	
6	*huragan*	
5	*nawałnica*	≈ phenomena related to a storm
5	*wiatr*	
1	*piorun*	
4	*wicher*	
4	*burzyć*	
3	*bunt*	
2	*salwa*	≈ in context, free association
2	*padać*	
2	*sztormowy*	
1	*zbuntowany*	

German equivalents for Czech *křičet*

1135	*schreien*	2	*losschreien*
185	*rufen*	2	*Rufen*
60	*brüllen*	1	*aufheulen*
31	*anschreien*	1	*dazwischenrief*
21	*Schrei*	1	*durcheinanderschrien*
16	*schreiend*	1	*greinte*
10	*kreischen*	1	*grölen*
6	*aufschreien*	1	*herriefen*
6	*Geschrei*	1	*herumzubrüllen*
3	*zurufen*	1	*hinausschreien*
3	*zuschrie*	1	*Lamentieren*
2	*anschreie*	1	*nachrufen*
2	*brüllend*	1	*quieken*
2	*geschrieen*	1	*schreit*
2	*herumschreien*	1	*schriest*
2	*hineinrufen*	1	*stöhnen*

Ver. 2.0

Source language: Czech

Target language: English

Restrict to: Collection(s): 6

prozvonit

Search

☑ Lemma ☐ Multiword ☐ RegEx ☑ A = a

Frequency	Proportion	Czech	English
3	25.0	prozvonit	ring
3	25.0	prozvonit	call
2	16.7	prozvonit	buzz
1	8.3	prozvonit	follow
1	8.3	prozvonit	again
1	8.3	prozvonit	over
1	8.3	prozvonit	bell
12			

Ver. 2.0

Source language: Czech

Target language: English

Restrict to: Collection(s): 6

hrotit

Search

☑ Lemma ☐ Multiword ☐ RegEx ☑ A = a

Frequency	Proportion	Czech	English
3	23.1	hrotit	push
2	15.4	hrotit	rise
1	7.7	hrotit	escalate
1	7.7	hrotit	sweat
1	7.7	hrotit	mess
1	7.7	hrotit	easy
1	7.7	hrotit	tense
1	7.7	hrotit	point
1	7.7	hrotit	grump
1	7.7	hrotit	dramatic
13			

Source language: Czech

Target language: English

Restrict to: Collection(s): 6

Ver. 2.0

hrotit

Search

☑ Lemma ☑ Multiword ☐ RegEx ☑ A = a

Frequency	Proportion	Czech	English
1	16.7	hrotit	to point
1	16.7	hrotit	push real hard
1	16.7	hrotit	easy
1	16.7	hrotit	to mess with
1	16.7	hrotit	push
1	16.7	hrotit	do dramatic
6			

▲ Frequency ▼	▲ Proportion ▼	▲ Polish ▼	▲ Czech ▼
56	20.9	kilkanaście	několik
19	7.1	kilkanaście	pár
10	3.7	kilkanaście	tucet
7	2.6	Kilkanaście	Několik
6	2.2	kilkanaście	několik desítek
5	1.9	kilkanaście	dvacet
4	1.5	kilkanaście	patnáct
4	1.5	kilkanaście	několika
3	1.1	kilkanaście	o pár
3	1.1	kilkanaście	deset
3	1.1	kilkanaście	dvanáct
3	1.1	kilkanaście	patnácti
3	1.1	kilkanaście	asi deset
3	1.1	kilkanaście	pár desítek
2	0.7	kilkanaście	jen několik
2	0.7	kilkanaście	asi tucet

Dissemination of texts

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

Teaching and research

- https://www.korpus.cz/biblio: 105 entries
- https://ukaz.cuni.cz/: 170 entries
- https://www.researchgate.net/: 49 entries
- https://scholar.google.cz/: 2090 entries


InterCorp
korpusu
of
the
and
ekvivalenty
parallel
českéčeštině
corpus
in
srovnání
The
konstrukce
Multilingual
překladech
parallelle
paralelním
Intercorp
Parallel
view
francouzštině
de
ich
jejího
en
českém
Dvojtečka
ruské
idiom
polskich
materiále
deutsch-tschechischen
preposition
podstawie
Corpus
Translations
point
et
corpora
equivalenten
hacer/fare
němčině
und
českých
paralelního
Czech
equivalents
češtiny
rrównoległego
InterCorpu
přeskladzie
sufixem
protějšky
současné
korpusů
open-choice
czeskich
multilingue
principle
Shifts
italské
francouzské
význam
vyjádření
překladu
angličtině
italštině
korpusová
sloves
Nederlandse
English
Anglické
nepředložkové
česko-ruského
korpus
Španělské
španělštině
překládání
analýza
základě
translation
Využití
Možnosti
předložek
from
mit
data
Case
Le
česko-francouzské
češtinou
paralelních
korpusie
německé
světle
v/ve
ekvivalentů
sloveso
between
francouzštiny
los
stany
czeskich
el

	áček	ánek	ásek	átko	eček	ečka	ečko	ek	enka	énko	íček	ička	ičko	íčko	ík	ínek	inka	ínko	ítko	ka	ko	oušek	uška	–	Σ
aczek	37		1	246	7	3		18			5	5		2	6	2	5			5	2	1		12	357
aszek	54				1			63			9	2			43				1	39		1		17	230
ątko		3		123	12			7	3		27	5		11	8		2			2		3		3	209
cia	4				2		4		22			570			1		6			64		1	1		675
eczek	39				56	1		100	1		83	41	1	8	68	1	1			6	16	10		60	492
eczka	59	2	1	12	24	55		48	11	5	112	367	1	5	39		20		1	253	13	5	20	200	1254
eczko	7				6		374	4		1	1	38		82		1	1			5	5			32	557
ek	176	127	218	30	173	5	8	1402	6		159	217	1	41	279	73	30		21	306	83	12		426	3796
ełko	4			14	1			1			4	197		23	1	4	4			5	102			4	364
enka								19	4			45			1		84			24	2				180
eńka									62								1			3					66
enko										141		2						3			54				200
iczek					9			27			125	31		10	1				4	2		1		13	223
iczka								29				355					1			28				6	419
ik	12		2	2	84	9		1167			75	132		18	160	1	23			92	25			201	2003
iszek											45				3					14				2	64
ka	103	26	3	209	70	56	7	568	171	5	66	1628	77	24	80	24	229	3	17	2166	43	11	48	357	6007
ko	2			119	14		69	278			14	19	10	211	16	2	10	10	9	4	244	2		52	1085
unia						1			13			145			1		1				1	2	1	1	166
uś	1			1	3			8			5	2			8	145	1			4	1			3	182
usia						1						14					137			6			2	1	161
uszek	5	1		1	130			109			16	73		1	183	4	12			5	2	34	4	21	601
uszka	1				1			5	55		60	16			10					4		2	5	1	160
yczek				1	1	1		61			22	2			22	27	3		3	3	1			10	157
yczka											2	140			4		1			6	1			2	156
yk	74	26		5	177	7		248			95	82		26	287	60	7	1		40	11			122	1268
–	590	474	144	751	1508	78	205	6051	146	202	1489	4540	99	769	1732	839	2131	111	664	4815	1130	235	62		28815
Σ	1187	664	369	1520	2310	218	680	10241	500	354	2436	8747	191	1266	2975	1199	2730	133	720	7949	1744	334	145	1559	50242

Object arguments of *toužit po* + N and pl equivalents

88	**chcieć**	71	**marzyć**	59	**pożądać**	216	**pragnąć**	107	**tęsknić**
22	ten	10	ten	13	on	41	ten	13	on
10	co	7	nic	7	žena	20	on	7	ten
6	on	4	co	4	já	15	co	5	který
2	jenž	4	který	4	který	12	ty	4	co
2	nic	3	jaký	4	ten	9	který	4	něco
2	voda	3	život	2	co	8	nic	4	ty
1	adopce	2	hodnost	2	druhý	5	já	3	den
1	blízkost	2	jenž	2	jenž	5	jenž	3	jenž
1	Catherine	2	něco	2	sláva	5	něco	3	láska
1	chvíle	2	sláva	2	věc	4	klid	2	bratrství
1	což	2	změna	1	barva	4	moc	2	čas
1	čest	1	cestování	1	děvče	4	smrt	2	já
1	čin	1	chalupa	1	jaký	3	dítě	2	jídlo
1	domácnost	1	chvíle	1	křeslo	3	jaký	1	bod
1	Estella	1	cizina	1	lovec	3	spravedlnost	1	byt
1	hádka	1	člověk	1	manželka	3	změna	1	cit
1	hodnost	1	dát	1	mír	2	Lucie	1	domov
1	kámen	1	den	1	ostrov	2	možný	1	domov
1	klid	1	divadlo	1	síla	2	odpověď	1	felicie
1	kompromis	1	experiment	1	slast	2	peníze	1	gesto
1	kontakt	1	jediný	1	stvoření	2	svět	1	Herder
1	který	1	katastrofa	1	tělo	2	svoboda	1	hlas
1	láska	1	krajka	1	trofej	2	tělo	1	Honorius
1	medvěd	1	krásný	1	úcta	2	útěk	1	jednota
1	místečko	1	letadlo	1	všechen	1	bohatství	1	klid
1	návrat	1	lezení	1	zvuk	1	cigareta	1	království
1	něco	1	maturita	1	ženství	1	cokoli	1	kultura

Most frequent object arguments of *toužit po*

	chcieć	marzyć	pożądać	pragnąć	tęsknić	Total	Grand Total
ten	22	10	4	41	7	84	100
on	6	1	13	20	13	53	69
co	10	4	2	15	4	35	46
který	1	4	4	9	5	23	30
nic	2	7		8	1	18	21
ty	1			12	4	17	17
jenž	2	2	2	5	3	14	22
něco	1	2		5	4	12	15
já			4	5	2	11	12
žena		1	7			8	8
jaký		**3**	1	**3**		7	8
klid	1			4	1	6	7
láska	1			1	3	5	8
den		1		1	3	5	6
smrt				4	1	5	6
tělo	1		1	2	1	5	5
změna		2		3		5	5
moc				4		4	8
sláva		**2**	**2**			4	6
svoboda		1		2	1	4	4
život		3		1		4	4
domov				1	2	3	4
dítě				3		3	3
hodnost	1	2				3	3
návrat	**1**	**1**			**1**	3	3
spravedlnost				3		3	3
věc			2	1		3	3
Total	50	46	42	153	56	347	426
Grand Total	87	71	59	216	107	541	675

Most frequent object arguments of *toužit* + inf

	chcieć	marzyć	pożądać	pragnąć	tęsknić	Total	Grand Total
být	8	6	1	26		41	44
vidět	4			7	1	12	16
jít	3			6		9	10
mít	2	2		4		8	11
poznat	1			5	2	8	10
vrátit	**3**	1		**3**	1	8	9
udělat	1			6		7	8
stát	3	2		1		6	8
zůstat	4			2		6	7
říci	4					4	5
spatřit				4		4	5
obejmout				4		4	4
slyšet	1	1		1	1	4	4
žít	2			2		4	4
setkat	1			2		3	5
dostat	1			2		3	4
milovat				3		3	4
najít	1			2		3	4
zbavit				3		3	4
dát	2			1		3	3
mluvit	1			2		3	3
potkat	1		1	1		3	3
promluvit	1	2				3	3
předvést	2			1		3	3
spasit	1			2		3	3
spát	1	2				3	3
změnit	1			2		3	3
znát	2			1		3	3
Total	51	16	2	93	5	167	193
Grand Total	116	35	4	197	16	368	446

Outline

Pre-processing

1. Scanning & character recognition
2. Proofreading
3. Segmentation (sentence boundary detection)
4. Alignment
5. Checking of segmentation and alignment
6. Tokenization, morphosyntactic markup

Tools used in pre-processing

1. Bibliographical database
2. *Intertext* – alignment editor
3. *Punkt* – sentence splitter
4. *Hunalign* – aligner
5. Language-specific tokenizers and taggers

Linguistic markup

= lemmatization and tagging

Strategy

- Use available tools (taggers), including:
  - Tokenization bundled with the tool
  - Tagsets designed elsewhere by experts on the given language
  - Annotation models trained elsewhere

Tools used for lemmatization and tagging

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

Outline

Access statistics

- September 2017 – August 2018
- Total no. of queries: 202 thousand $\approx$ 553 per day
- For each query the first language and any other language
- Queries into custom corpora not included
- About 2.8% queries select 3 or more languages (2015)

Queries by L1 and L2

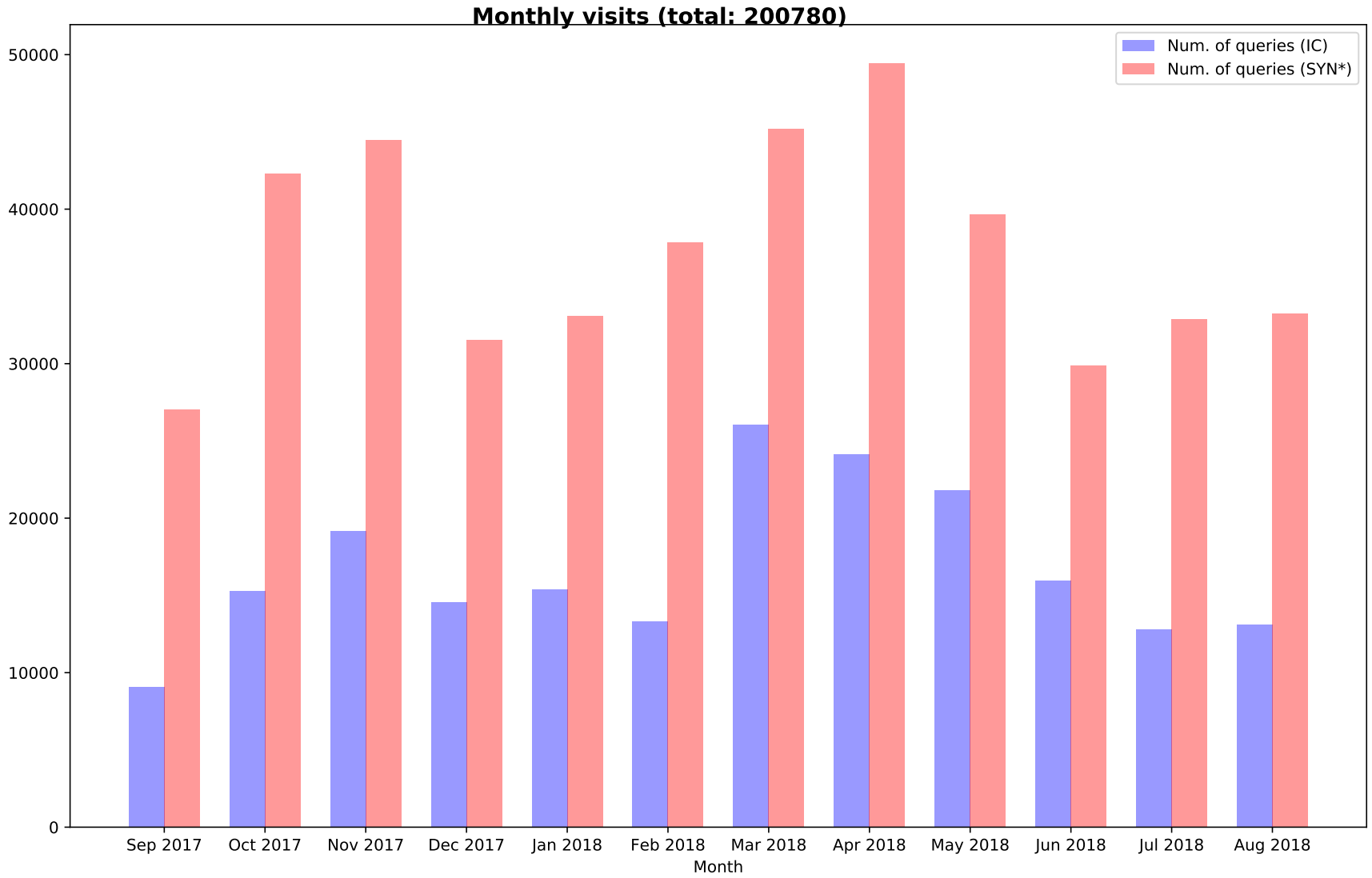
- L1: **cs** *63k*, **en** *42k*, **de** *38k*, **fr** *15k*, **es** *12k*
  ... **ru** *6.2k*, **pl** *4.0k*, **bg** *3.1k*
- L2: **cs** *98k*, **en** *29k*, **none** *27k*, **de** *18k*
  ... **ru** *4.9k*, **pl** *3.2k*, **sk** *1.5k*, **hr** *1.1k*
- Combinations:
  en→cs *33k*, **de→cs** *25k*, **cs→en** *24k*, **cs→de** *16k*
  ... **ru→cs** *4.1k*, **bg→cs** *3.0k*, **cs→ru** *2.2k*, **pl→cs** *2.1k*, **cs→pl** *1.6k*

	bg	cs	de	en	es	fi	fr	hr	hu	it	lt	nl	pl	ru	sk	sl	sr	sv	uk	none	Σ
bg	0	2966	6	2	0	0	0	2	1	0	0	0	8	24	4	3	2	0	0	55	3073
cs	435	0	16266	24248	3970	711	5427	644	81	904	524	556	1638	2246	201	69	164	670	217	3996	62967
de	14	25369	0	1106	91	7	176	252	67	312	0	39	390	1081	212	12	32	14	8	8405	37587
en	11	32852	447	0	148	276	182	25	12	2274	15	30	196	169	264	5	2	44	7	4784	41743
es	2	6933	100	642	0	0	31	0	0	177	0	1	21	202	6	0	0	1	0	3507	11623
fi	0	4729	70	127	0	0	13	0	0	16	1	0	553	0	687	0	0	12	0	1202	7410
fr	0	11860	148	333	181	3	0	0	0	155	0	0	6	25	2	0	0	1	0	1789	14503
hr	14	213	17	8	0	0	0	0	0	0	9	0	77	32	12	9	2	0	9	185	587
hu	0	90	19	2	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	124	238
it	0	1722	132	1107	747	6	63	0	0	0	0	11	0	383	32	0	0	7	0	1503	5713
lt	0	549	0	22	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	581
nl	0	524	15	9	106	0	1	0	0	1	0	0	0	8	0	0	0	0	2	212	878
pl	24	2082	174	267	4	172	1	75	0	6	0	0	0	544	17	14	10	9	33	532	3964
ru	98	4086	304	371	117	16	20	117	0	110	2	9	235	0	37	70	62	54	60	454	6222
sk	0	145	73	326	9	8	16	16	13	22	0	0	15	27	0	0	0	5	17	83	775
sl	3	38	1	1	0	0	0	0	0	0	0	0	54	51	0	0	0	0	0	43	191
sr	0	49	20	3	0	0	0	0	0	0	0	0	0	63	0	4	0	0	0	44	183
sv	0	3313	7	21	0	0	1	0	2	0	0	0	0	4	3	0	0	0	0	118	3469
uk	0	18	0	3	0	0	0	1	0	0	0	0	1	36	0	1	0	0	0	77	137
Σ	601	97538	17799	28598	5373	1200	5932	1132	176	3977	551	646	3194	4897	1477	187	274	817	353	27122	201844

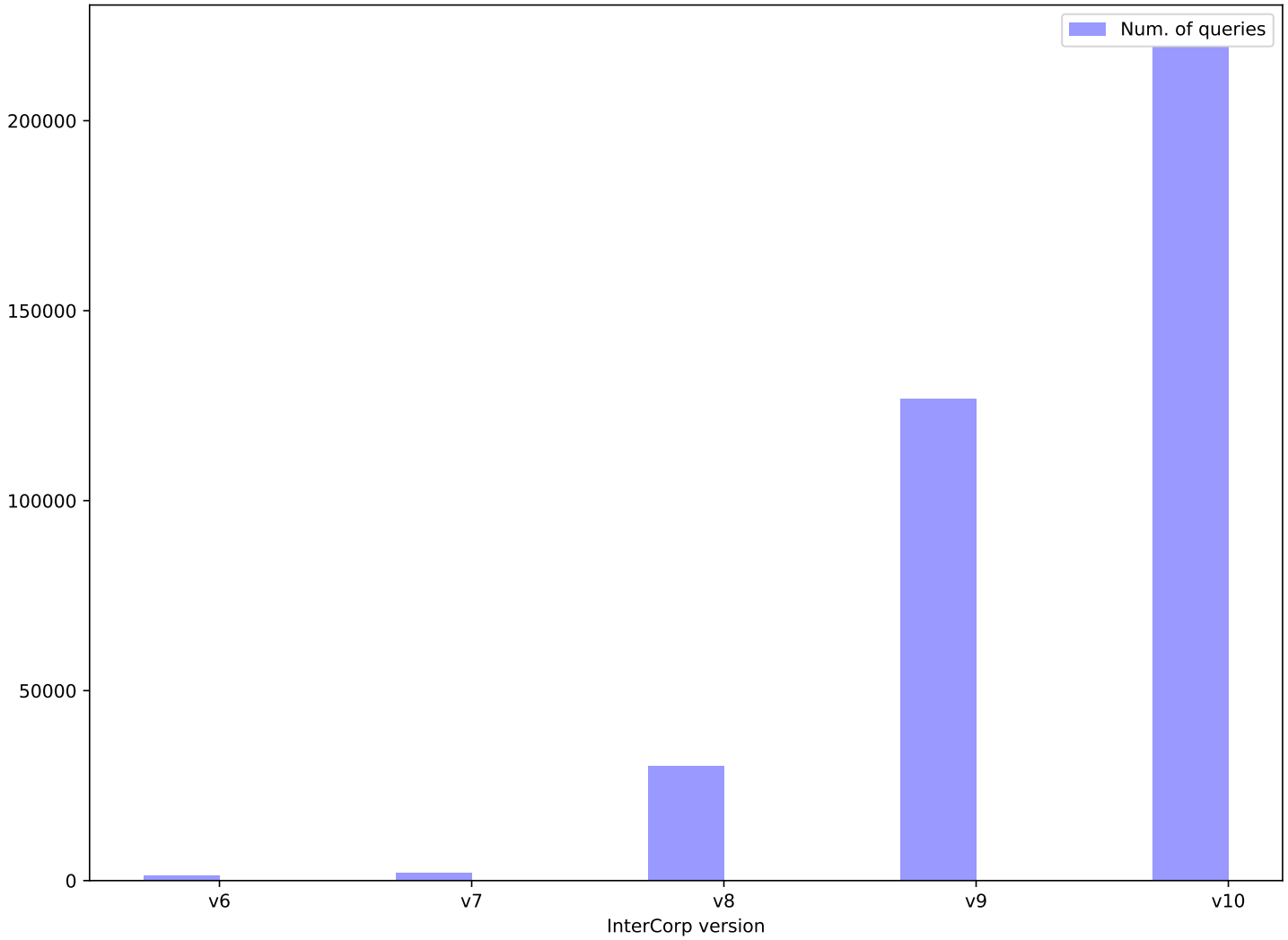
Queries by text groups

ALL	184,586	88.9%
Core	11,053	5.3%
Law	1,875	0.9%
Journalism	1,411	0.7%
Subtitles	1,061	0.5%
Proceedings	807	0.4%
Bible	347	0.2%
Total	**207,680**	**100.0%**

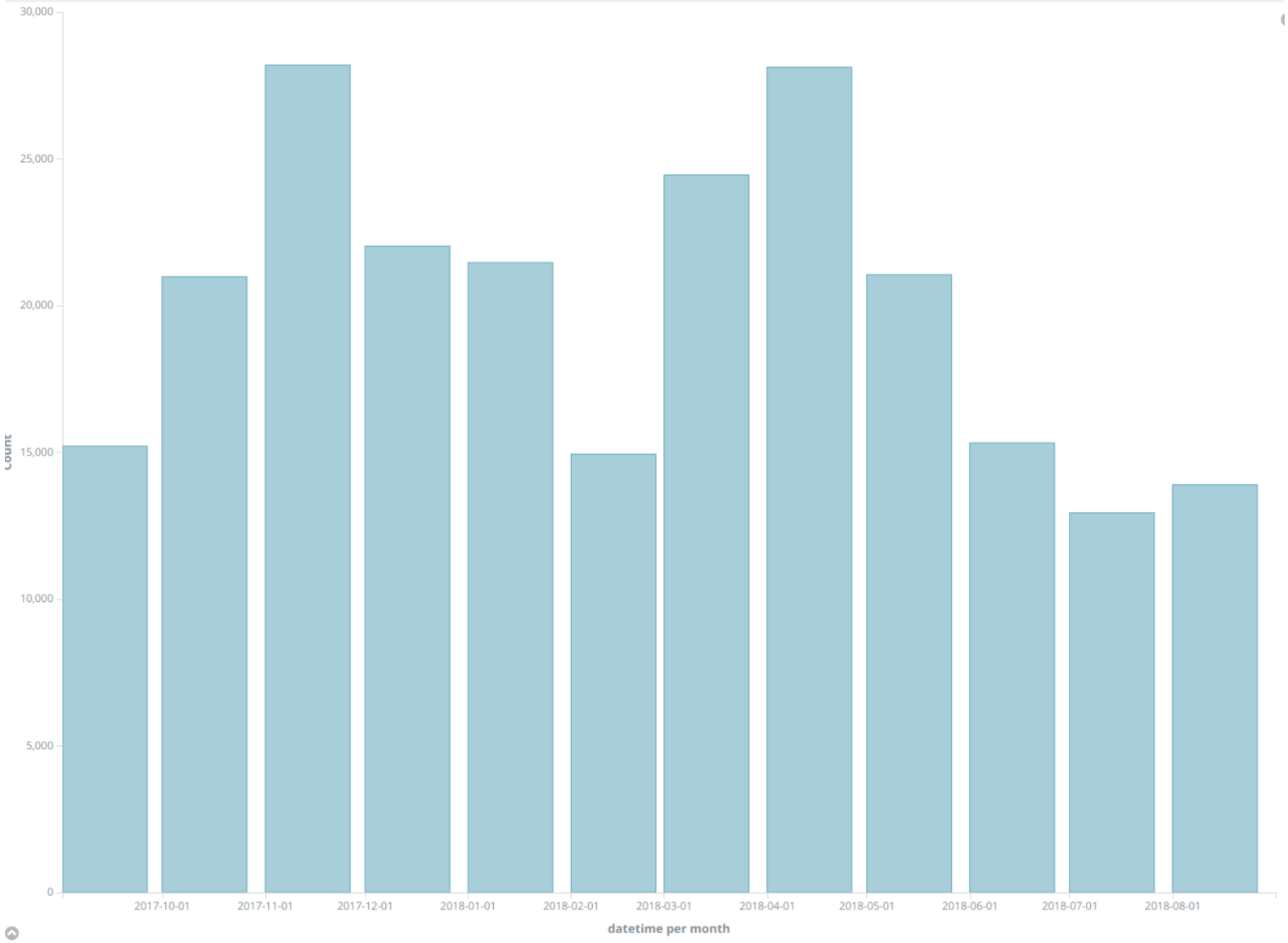
InterCorp vs. SYN*



InterCorp releases



Treq


30,000
25,000
20,000
15,000
10,000
5,000
0
Count
2017-10-01
2017-11-01
2017-12-01
2018-01-01
2018-02-01
2018-03-01
2018-04-01
2018-05-01
2018-06-01
2018-07-01
2018-08-01
datetime per month

Survey (winter 2015/2016)

- **Asked** 748 users, response rate 17.4% (130)
- Most popular **text types**: core (62%), whole corpus (42%), Syndicate (17%), Europarl (14%), Acquis (13%), VoxEurop (12%), Subtitles (8%)
- **Motivation**: contrastive analysis (74%), equivalents (69%), translatological analysis (61%), single language (27%), FLT preparation (22%), FLT in-class (21%), NLP (4%)
- **Desiderata**: larger core (55 %), more translations in one language (52%), other text types (42%), inclusion of the original (41%), as many translations of a text as possible (31%), tagset harmonization (30%), balanced subcorpora (25%), other languages (7%)
- **Positives**: choice of languages, correct alignment, partitioning and custom corpora, extensions and improvements, free access, user support, availability
- **Negatives**: size, translation quality (Subtitles), disparate tagets, missing metadata, technical drawbacks of the interface

Outline

Content

- 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

Search interface

- Missing functionalities:
  - **biKWiC** – highlighting keyword equivalent
  - Info on **alignment**: 1:1 / 2:1 / 1:2 / automatic / manual / confidence score
- Tools beyond mere search
  - **Comparisons** across text types, languages, corpora ...
  - **Co-occurence** profiles [Belica(2011)]
  - Word **sketches** [Kilgarriff et al.(2014)]

Annotation

- Languages differ in tagsets and tokenization rules
  – **harmonization of tagsets**
- Morphosyntactic annotation **for all languages**
- **Alignment** by words, multiword units, constituents
- **Syntactic** annotation
- **Crowdsourcing** to eliminate annotation errors

Tokenization

- *abychom, udělals, tys, očs, zum, aux*
  $\times$
  že by śmy, zrobił eś, ty ś, gdzieś/gdzie ś, ca n't, I 'm
- *Estados~Unidos* (NP), *a~lo~largo~de* (PREP), *tendrán~que* (VMfin), *por~el~momento* (ADV), *al~mismo~tiempo* (ADV)
- *cure-dents, gut-ausgearbeitet, Jelzin-Ära, franco-tedesco*
  $\times$
  padne - li, Tchaj - wan, česko - německý

Tagsets

- *under*, *because*: en:**IN**
- *těch*: cs:**PD** $\times$ *tych*: pl:**adj**
- *devátá*: cs:**Cr** $\times$ *dziewiąta*: pl:**adj**
- *remotest*: en:**JJS** $\times$ **abgelegenste**: de:**ADJA**

Solutions?

Universal Dependencies

- http://universaldependencies.org/
- A de-facto standard also for morphological categories
- Loss conversion from language-specific tagsets
- Incompatible tokenization

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 while 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

Alternatives to UD?

- Lossless mapping of tagsets onto an ontology of linguistic categories [Chiarcos & Erjavec(2011)], [Chiarcos(2012)]
- Multidimensional taxonomy of morphosyntactic categories

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, other as complementary
  - Semantics (Jespersen, [Brøndal(1928)], Vinogradov, Tesnière,...)
  - Morphology[1] ([Saloni & Świdziński(1985), 95])
  - Syntax ([Grzegorczykowa et al.(1998), 59])
  - Syntax/Morphology ([Komárek et al.(1986), 13–16])

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

Parallel corpus in 3D

- Cross-classification to embrace existing tagsets:
  - Czech (ÚFAL): preference for **semantic** classes
  - Polish (IPI PAN): **inflectional** classes
  - German (STTS): preference for **syntactic** classes

Linking parallel corpora

- Parallel corpora are useful to speakers of any language
- High synergy in infrastructure and content:
  - Many problems are similar across languages
  - Texts in foreign languages may exist elsewhere
  - Native speakers are the best corpus builders
- Options / Levels of cooperation:
  - Exchange of know-how, tools, texts between centres
  - Virtual integration of content, a common search interface (federated search), a common text dissemination policy
  - A single centre providing coordination and infrastructure for all languages


vám
Благодаря
Спасибо
wam
Дякую
Ви
Ďakujem
Hvala
vse
je
благодарам
Дзякуй
Dziękuję
lijepa
Děkuji

Outline

Belica, C. (2011).
Semantische Nähe als Ähnlichkeit von Kookurenzprofilen.
In A. Abel and R. Zanin, editors, *Korpusinstrumente in Lehre und Forschung*, pages 155–178, Brixen. Bozen-Bolzano University Press.

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.

Kaczmarska, E. & Rosen, A. (2016).
Syntakticko-sémantický popis vybraných skupin sloves vyjadřujících emoce a pocity – metody kontrastivního zkoumání valence na základě paralelního korpusu.
In K. Skwarska and E. Kaczmarska, editors, *Výzkum slovesné valence ve slovanských zemích*, volume 43 of *Práce Slovanského*

ústavu AV ČR – Nová řada, pages 319–350. Slovanský ústav AV ČR, Praha.

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., Kaczmarska, E., & Škodová, S. (2014).
Zdrobnienia jako element kultury i pułapka glottodydaktyczna. Czeskie i polskie deminutiva w ujęciu konfrontatywnym na podstawie badań korpusowych [Diminutives as a cultural element and a glottodidactic trap – Czech and Polish diminutives from a contrastive corpus-based perspective].

In E. Kaczmarska and A. Zieniewicz, editors, *Glottodydaktyka wobec wielokulturowości*, pages 51–66, Warszawa. Wydział Polonistyki Uniwersytetu Warszawskiego.

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