Paralelní korpusy

Korpusová a teoretická lingvistika
Teoreticko-metodologický seminář
10. prosince 2020

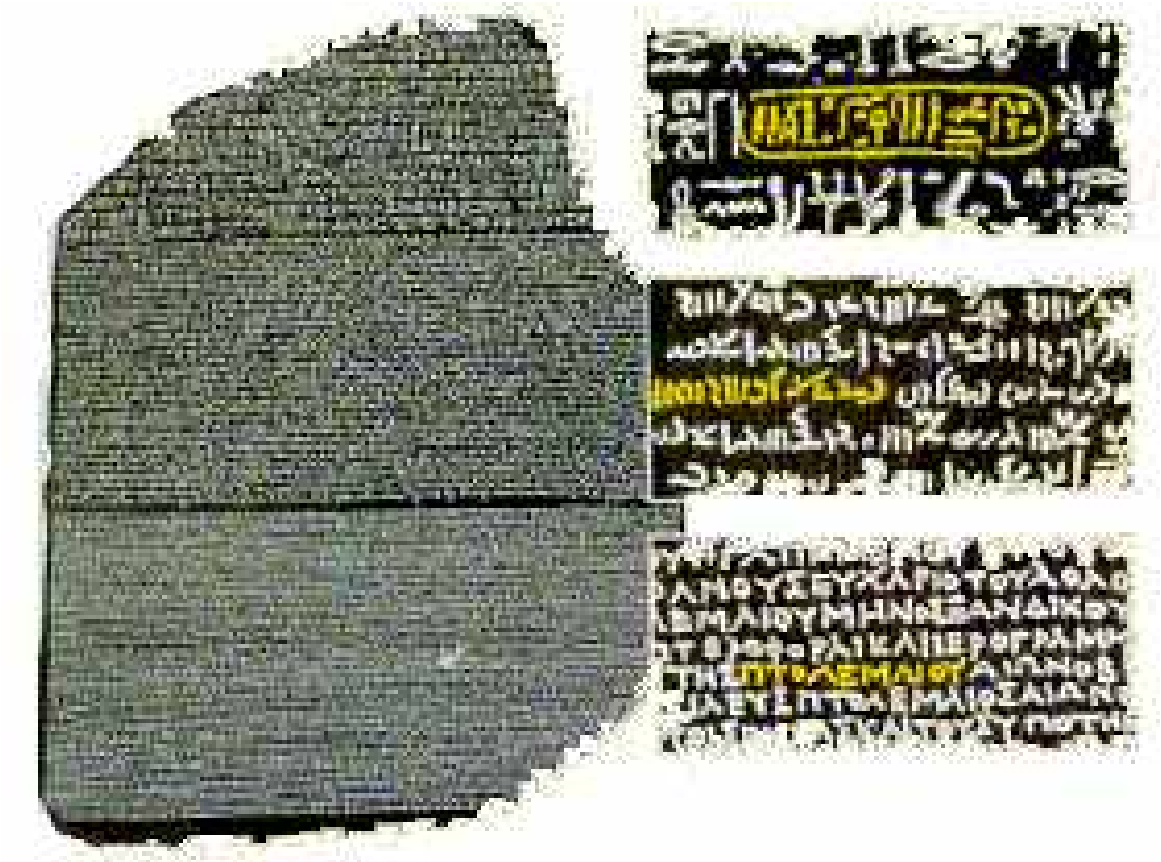
Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Filozofické fakulty Univerzity Karlovy v Praze
alexandr.rosen@ff.cuni.cz
http://utkl.ff.cuni.cz/~rosen
http://utkl.ff.cuni.cz/~rosen/public/pc2020.pdf

- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Co je to paralelní korpus?

- Paralelní korpus obsahuje stejné nebo srovnatelné texty ve více podobách, které se liší jazykem, verzí překladu, dialektem, ...



- Stejný text ve **více verzích** (jazycích, překladech, ...)
- Z **paralelních textů**, přeložených nebo rovnou vytvořených ve více jazycích
- Paralelní korpusy (i když se jim tak neříká) **využívají**:
  - **nástroje** pro zpracování přirozeného jazyka: strojový překlad, vyhledávání informací, projekce anotací, ...
  - **překladatelé** (Computer-Assisted Translation)
  - experti na **výuku cizích jazyků**
  - **lexikografové**
  - **translatologové**
  - ...

Typy paralelních korpusů:

- srovnatelné (texty ze stejného oboru, nikoli překlady)
- překladové

Většinou se *paralelní* korpusy ztotožňují s *překladovými*.

Další faktory

- velikost
- jazyky
- zarovnání
- anotace
- typy textů
- dostupnost

Význam a překlad

- Překlad zachovává význam
- Paralelní kontext
  - → explicitní překladová ekvivalence
  - → implicitní anotace významu
- Od významu k formě:
  - najdi ekvivalenty v jiném nebo stejném jazyku
  - – translatologie, kontrastivní lingvistika, výuka cizích jazyků, strojový překlad, překlad s pomocí počítače (CAT)
- Od formy k významu:
  - najdi význam prostřednictvím jiných jazyků
  - – porozumění textu, projekce anotace, jednojazyková lexikografie

Předpoklady pro rozumnou práci s paralelními korpusy:

- Zarovnání po úsecích textu (textech, odstavcích, **větách**, větných členech, slovech)
- Paralelní korpusový manažer (*concordancer*)

Problémy

- Autentičnost
  - *translationese*
- Dostupnost
  - ne ve všech jazycích, žánrech, typech textů
  - právní omezení
- Zarovnání
  - není bez chyb
- Jsou potřeba zvláštní nástroje
  - pro zarovnávání
  - pro paralelní vyhledávání

- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Rovnou pro lidi:

- Pro překladatele
  - paralelní konkordance
  - překladová paměť
    (*Translation Memory*, v programech pro podporu překladu)
- Pro učitele a studenty cizích jazyků
- Pro lexikografy
  - paralelní konkordance
  - extrakce ekvivalentů slov nebo kolokací
- Pro translatology, literární vědce, komparatisty, dialektology
- Pro ostatní lingvisty taky!

Pro aplikace:

- Strojový překlad (*Machine Translation*)
  - pomocí neuronových sítí (*Neural Machine Translation*) [Koehn(2020)]
  - statistický (*Statistical Machine Translation*) [Bojar(2012)]
  - podle příkladů (*Example-based Machine Translation*)
  - hybridní (*Hybrid Machine Translation*)
- Vyhledávání informací ve více jazycích (*cross-language information retrieval*)
- Projekce anotace (interpretace textu v jednom jazyce na základě jazyka druhého)

- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

determined I

determined II

Ve slovníku (Hais – Hodek, Academia 1991):

determined

1. rozhodný, zarytý
2. rozhodnutý, odhodlaný, zamanuvší
3. v. *determine*

determine

1. určit, určovat, stanovit, udat, udávat
2. rozhodnout, učinit rozhodnutí
3. rozhodnout se
4. zjistit, vyšetřit, vypočíst
5. přimět
6. zanikat, končit, ukončit
7. vymezit, ohraničit

determined III

By now Les had engineered dozens of multiple-recorded discs and was determined that the world hear them.
Hackman returned to New York determined to succeed.
But Mr. Hill certainly had it, and I was determined to see how it worked.
Steven was determined to make himself understood.
Now, however, as the trial progressed, Donna grew stronger and more determined.
Kallie rose slowly, determined to please her mistress.
But that only made me more determined.

Les měl tou dobou už desítky vícestopě nahraných desek a usiloval o to, aby je uslyšel i svět.
Hackman se vrátil do New Yorku s předsevzetím, že prorazí.
Pan Hill ji však zcela jistě vzbuzoval a já chtěl vidět, jak toho dociluje.
Steven měl všechny předpoklady pro to, aby se naučil mluvit.
Jak se však proces vyvíjel, Donna se zocelovala a odhodlávala.

Kallie se zvedala pomalu, ale s odhodláním potěšit svou paní.
Tím však jen posílili mé odhodlání.

determined IV

When a reunion of the Point Cruz crew was organized for September 1993, Bill was determined to have “George” there.
As a young factory worker, Sheets was determined to give his three children summers they would always remember.
Eager to impress the head keeper with my animal-handling expertise, I made a determined grab.

If you find yourself going flat or tentative, determined thoughts can make all the difference.

Když se bývalí členové posádky dohodli, že se v září 1993 zase po letech sejdou, zařekl se Bill, že tam "George"nesmí chybět.
Když ještě zamlada pracoval v továrně, umínil si, že svým třem dětem dopřeje letní prázdniny, na jaké nikdy nezapomenou.
Ale já jsem chtěl hlavního ošetřovatele ohromit svou zručností při manipulaci se zvířaty a rázně jsem bažanta popadl.
Když se vám zdá, že ochabujete nebo že se cítíte nejistí, vše můžou napravit pevné, vyhraněné myšlenky.

determined V

Even before the diagnosis was confirmed, the Odones, both determined, strong-willed people, had decided they would learn all they could about the disease.

Ještě před potvrzením diagnózy se Odoneovi, oba cílevědomí a nezdolní lidé, rozhodli, že si o té chorobě zjistí, co se dá.

I would close my eyes, determined not to give him the satisfaction of seeing me cry.

Jen mu neudělat radost, jen se nerozbrečet!

sophisticated I

Ve slovníku (Hais – Hodek, Academia 1991):

sophisticated

1. příliš zkušený, znalý světa, blazeovaný, náročný, intelektuálně na výši, vysoce kultivovaný, překultivovaný
2. výlučný, exkluzivní, vysoce náročný, pro úzký okruh
3. (stroj) velmi složitý, komplikovaný, (zbraň) sofistikovaný; (teorie) složitý, subtilní, rafinovaný, vyspekulovaný
4. (auto) s posledními technickými vymoženostmi
5. klamný
6. viz *sophisticate, v.*

sophisticated II

This led to the development of synchronized stereophonic tape, right up to the sophisticated present.
This technological marvel has become amazingly sophisticated.

At the city's Wat Nai Rong High School, 17-year-old Wasana Warathongchai says smoking makes her feel "sophisticated and cosmopolitan, like America."
I didn't get a buzz, because I didn't inhale, but just the fact I was actually smoking made me think I was cool sophisticated.

To vedlo k vývoji synchronizované stereofonní nahrávky v její dnešní dokonalosti.
Tato technická hříčka se totiž v poslední době podivuhodně zdokonalila.
Sedmnáctiletá studentka střední školy Wasana Warathongchai vysvětluje, že když kouří, „připadá si moderní a kosmopolitní jako Amerika."
Nic to se mnou neudělalo, protože jsem nešlukovala, ale pocit, že doopravdy kouřím, byl fantastický.

sophisticated III

Kids or teen-agers who think smoking is cool sophisticated or who want to try it: don't!
Today, after years of research, educators are more sophisticated about detecting learning disabilities and teaching children how to compensate for them.
Scientists had processed the images and additional ones from sophisticated Landsat satellites, which used a number of light and radio wavelengths to detect surface details.
I wanted my mother to be more sophisticated, like my friends' mothers.

Všem klukům a holkám, kterým kouření připadá takové dospělé a rádi by to zkusili taky, chci říct: Nedělejte to!
Dnes, po mnohaletých výzkumech, jsou učitelé o poruchách schopnosti učení více informováni, umí je rozpoznat a vědí, jak takové děti učit.

Odborníci analyzovali snímky z vesmíru i fotografie získané z družic Landsat, které k mapování povrchu Země využívají světelné a radiové vlny.
Chtěla jsem, aby moje matka byla elegantní jako matky mých kamarádek.

sophisticated IV

And perhaps because, at still another level, we enjoy watching their gloriously sophisticated competition for our favors.
Fleming secured sophisticated radio pagers that would keep the surveillance teams in constant contact with the Bexleyheath control center and alert them if the Ian and Nina Fox cash card was being used at an ATM machine.
In the near future, data collection will become even more sophisticated.

Možná i proto, že na ještě jiné úrovni zálibně pozorujeme, jak rafinovaně se ucházejí o naši přízeň.

Fleming opatřil výkonná radiofonická pojítka, která umožňovala, aby sledovací týmy byly v nepřetržitém kontaktu s řídícím střediskem v Bexleyheathu a mohly je okamžitě uvědomit, kdyby někdo použil platební kartu Foxových.
V blízké budoucnosti se sběr dat v supermarketech stane ještě významnější disciplínou.

- Přehled
- Jak si vyrobit vlastní paralelní korpus

pozor, ne všecko je aktuální!

název	typy	jazyky	velikost	anotace	zarovnání	korektura	hledání	stažení	metadata
Linguee	právo	25	?	ne	V,S	ne	ano	ne	ano
Glosbe	různé	100+	1Bs	ne	V,S	ne	ano	ne	ano
SKE	různé	38	217M cs	ne	V	ne	ano	ano	ano
DGT-TM	právo	22	100M cs	M,Sy	V	ano	ne	ano	ne
Pelcra	různé	31	58M pl	ne	V,S	část	ne	ano	ano
RNC	různé	6	9M	M	V	část	ano	?	ano
SNK	belet.	7	388M sk	M	V	ne	ano	část	ano
CzEng	různé	en, cs	233M en	M,Sy	V	ne	ano	ano	ne
PCEDT	žurn.	en, cs	1.2M.	M,Sy,Se	V,S	ano	ano	ano	ano
Kačenka	belet.	en, cs	3.3M	ne	S	ano	ne	ano	ano
Opus	různé	100+	4.7B	M,Sy	V,S	ne	ano	ano	ne
Parasol	belet.	31	27M	M	V	část	ano	?	ano
ASPAC	belet.	25	68t	ne	O	ano	ne	?	ano
InterCorp	různé	41	1.8B	M	V	část	ano	ano	ano

- **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 – https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory, https://lindat.mff.cuni.cz/services/kontext/
- **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, https://lindat.mff.cuni.cz/services/kontext/
- **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

Další korpusy

- https://www.clarin.eu/resource-families/parallel-corpora
- https://lindat.mff.cuni.cz/services/kontext/corpora/corplist

OPUS – an open source parallel corpus

http://opus.lingfil.uu.se

- Evropská centrální banka *(19 jazyků, č.: 1,4 mil. vět, 29,3 mil. slov)*
- EU Bookshop *(48 jazyků, č.: 1 mil. vět, 16,3 mil. slov)*
- Evropská ústava *(21 jazyků, č.: 11 tis. vět, 128 tis. slov)*
- jednání Evropského parlamentu *(21 jazyků, č.: 669 tis. vět, 13 mil. slov)*
- systémová hlášení KDE *(92 jazyků, č.: 134 tis. vět, 696 tis. slov)*
- manuály PHP *(22 jazyků, č.: 63 tis. vět, 147 tis. slov)*
- dokumenty Evropské agentury pro léčiva (EMEA)
  (22 jazyků, č.: 1,2 mil. vět, 14,2 mil. slov)
- filmové titulky *(30 jazyků, č.: 1,8 mil. vět, 11,2 mil. slov)*

- Kačenka: Korpus anglicko-český Katedry anglistiky FF MU Brno, *celkem přes 3 mil. slov* http://www.phil.muni.cz/angl/kacenka/kachna.html
- PCEDT: Prague Czech-English Dependency Treebank http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.htm
  - Wall Street Journal *22k vět, 488k slov* – syntax
  - Reader's Digest *44k vět a 660k slov* – jen text
- Multext/East: 1984 *(George Orwell)* nl.ijs.si/ME/
- Acquis Communautaire:
  21 jazyků, č.: *6 mil. slov* http://wt.jrc.it/lt/Acquis/
- Parallel Corpus of Computer Terms – Slovenský národný korpus
  http://korpus.juls.savba.sk/pcct/index.sk.html
- CzEng: Czech-English Parallel Corpus, syntakticky anotovaný [Bojar & Žabokrtský(2009)] http://ufal.mff.cuni.cz/czeng10/
  - zákony EU, projekt Navajo, technická dokumentace, beletrie, zprávy, webové stránky, filmové titulky
    (č.: 15 mil. vět, 206 mil. slov)

ASPAC – the Amsterdam Slavic Parallel Corpus

- autor: Adrie Barentsen
- *InterCorp* ho obsahuje téměř celý
- celková velikost >4 mil. tokenů
- 49 textů alespoň ve 4 slovanských jazycích
- 10 textů alespoň v 10 různých slovanských jazycích
- 11 slovanských jazyků má aspoň 15 textů
- některé překlady jsou ve více verzích
  (6 ruských a 4 polské překlady *Alenky v říši divů*)
- obsahuje také horní a dolní lužickou srbštinu

ParaSol: A Parallel Corpus of Slavic and other languages

- autoři: Ruprecht von Waldenfels a Roland Meyer
- on-line na adrese http://parasol.unibe.ch
- 18 mil. tokenů (slovanské jazyky) + 7,6 mil. (ostatní)
- ruština: 3,6 mil. tokenů, polština 3,4 mil. tokenů
- většina jazyků je vybavena morfologickou anotací a lemmaty

- Přehled
- Jak si vyrobit vlastní paralelní korpus

Nástroje na přípravu a/nebo prohledávání paralelních korpusů

- InterText https://wanthalf.saga.cz/intertext
- Sketch Engine https://app.sketchengine.eu
- AntPConc http://www.laurenceanthony.net/software/antpconc/
- ParaConc https://paraconc.com

Jak reprezentovat informace o zarovnání

Podobně jako jiné typy anotace, může být online nebo standoff:

Inline:

- tabulka nebo soubor s oddělenými paralelními strukturami
- dva soubory s paralelně vyznačenými strukturami
- XML podle standardu Translation Memory eXchange (TMX)

Standoff:

- odkazy na struktury v paralelních textech
- odkazy na řádky v paralelních textech, tzv. žebřík (*ladder*)

Sketch Engine (inline)

Corpus 1

1	<s>
2	<align>
3	This
4	is
5	the
6	first
7	sentence
8	in
9	corpus
10	1
11	.
12	</align>
13	</s>
14	<s>
15	<align>
16	This
17	is
18	the
19	second
20	sentence
21	in

Corpus 2

1	<s>
2	<align>
3	This
4	is
5	the
6	first
7	sentence
8	in
9	corpus
10	2
11	.
12	</align>
13	</s>
14	<s>
15	<align>
16	This
17	is
18	the
19	second
20	sentence
21	in

TMX (inline)

```
<tmx version="1.4">
  <header
    creationtool="XYZTool" creationtoolversion="1.01-023"
    datatype="PlainText" segtype="sentence"
    adminlang="en-us" srclang="en"
    o-tmf="ABCTransMem"/>
  <body>
    <tu>
      <tuv xml:lang="en">
        <seg>Hello world!</seg>
      </tuv>
      <tuv xml:lang="fr">
        <seg>Bonjour tout le monde!</seg>
      </tuv>
    </tu>
  </body>
</tmx>
```

InterCorp (standoff)

```
<?xml version='1.0' encoding='utf-8'?>
<doc id="Patterson-President" language="en" versio
<p id="1">
<s id="1:1">
<w lemma="the" tag="DT">The</w>
<w lemma="president" tag="NN">president</w>
<w lemma="be" tag="VBZ">is</w>
<w lemma="miss" tag="VBG">missing</w>
</s>
</p>
<p id="2">
<s id="2:1">
<w lemma="Bill" tag="NP">BILL</w>
<w lemma="Clinton" tag="NP">CLINTON</w>
<w lemma="and" tag="CC">and</w>
<w lemma="James" tag="NP">JAMES</w>
<w lemma="Patterson" tag="NP">PATTERSON</w>
</s>
</p>
<p id="3">
<s id="3:1">
<w lemma="special" tag="JJ">Special</w>
<w lemma="thanks" tag="NNS">thanks</w>
<w lemma="to" tag="TO">to</w>
<w lemma="Robert" tag="NP">Robert</w>
<w lemma="Barnett" tag="NP">Barnett</w>
U:**-  Patterson-President.en-00.tag.xml   Top (21,1
```

```
<?xml version='1.0' encoding='utf-8'?>
<doc id="Patterson-President" language="cs" version="00">
<p id="1">
<s id="1:1">
<w lemma="pohřešovat" tag="VB-S---3P-AA---I">POHŘEŠUJE</w>
<w lemma="se" tag="P7--4----------">SE</w>
<w lemma="prezident" tag="NNMS1-----A-----">PREZIDENT</w>
</s>
</p>
<p id="2">
<s id="2:1">
<w lemma="Bill" tag="NNMS1-----A-----">BILL</w>
<w lemma="Clinton" tag="NNMS1-----A-----">CLINTON</w>
</s>
</p>
<p id="3">
<s id="3:1">
<w lemma="James" tag="NNMS1-----A-----">JAMES</w>
<w lemma="Patterson" tag="NNMS1-----A-----">PATTERSON</w>
</s>
</p>
<p id="4">
<s id="4:1">
<w lemma="zvláštní" tag="AAIP1----1A-----">Zvláštní</w>
<w lemma="dík" tag="NNIP1-----A-----">díky</w>
<w lemma="náležet" tag="VB-P---3P-AA---I">náleží</w>
U:**-  Patterson-President.cs-00.tag.xml   Top (23,12)   (nXML Invalid Spc F
Beginning of buffer
```

```
<?xml version='1.0' encoding='utf-8'?>
<linkGrp toDoc='Patterson-President.en-00.xml' fromDoc='Patters...
on-President.cs-00.xml'>
<link type='1-1' xtargets='1:1;1:1' status='auto'/>
<link type='1-2' xtargets='2:1;2:1 3:1' status='auto'/>
<link type='1-1' xtargets='3:1;4:1' status='auto'/>
```

Manatee (žebřík – standoff)

en	cs
0	0
1	1,3
-1	4
2,4	5
5	-1
6:8	6:8

- 1. anglická věta (0) = 1. česká věta (0)
- 2. anglická věta (1) = 2. až 4. česká věta (1,3)
- 5. česká věta (4) ze zarovnání vypadne (-1)
- 3. až 5. anglická věta = 6. česká věta
- 7. až 9. anglická věta = po řadě 7. až 9. česká věta
- v zarovnání musí být uvedeny všechny věty

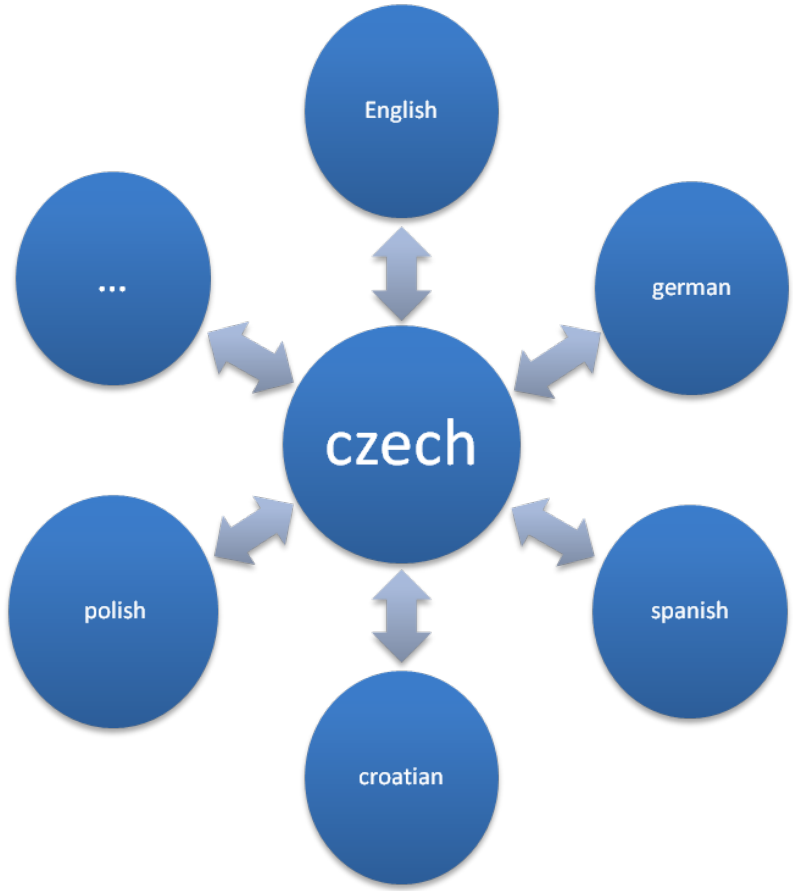
- Základní údaje
- Obsah korpusu
- Jádro
- Filmové titulky

Základní údaje

- *InterCorp* – vícejazykový paralelní korpus zaměřený na češtinu
- součást *Českého národního korpusu*
- https://intercorp.korpus.cz
- *2005
- zpočátku jako služba pro lingvistická pracoviště FF UK
- od 2008 on-line
- každý rok nové vydání

Architektura korpusu *InterCorp*

- zarovnání: po větách, údaje o zarovnání oddělené od vlastního textu
- každý text je česky a aspoň v jednom dalším jazyce
- zarovnání mezi texty v cizích jazycích přes českou verzi
- morfologické značky a lemmata pro většinu jazyků



Čím se InterCorp liší od jiných paralelních korpusů

- Velký podíl beletrie
- Korektury
- Bohatá metadata
- Stejné vyhledávací rozhraní jako ostatní korpusy ČNK
- Uživatelé se podílejí na tvorbě korpusu

Kritéria pro výběr textů

- Text se dá nějak získat
- Kvalita předlohy (souboru) dostatečná
- Text je:
  - úplný
  - jeho členění odpovídá jiným verzím
  - překlad je dobrý
- Typ textu:
  - reprezentativnost
  - vyvážení skladby korpusu
- Stejný text už je v jiných jazycích
- Jde o
  - originál,
  - překlad už existujícího českého originálu nebo
  - český překlad

Kdo je za co odpovědný

- Ústav Českého národního korpusu:
  - management, finance
  - technická podpora, školení, konzultace
  - centrální datové úložiště
  - formátování textů, dělení vět
  - automatické zarovnání, morfosyntaktické značkování a lemmatizace
- Koordinátor pro daný jazyk:
  - výběr a akvizice textů
  - korektury textů a zarovnání

Spolupráce

- Získávání a příprava textů:
  - Univerzita Karlova v Praze
  - Masarykova Univerzita v Brně
  - Univerzita Palackého v Olomouci
  - Česká akademie věd
  - Varšavská univerzita
- Pomoc ze zahraničí:
  - texty (ASPAC, Parasol, OPUS, ...)
  - nástroje pro lingvistickou anotaci (TreeTagger, ...)
  - obecnější nástroje pro zpracování textu (HunAlign, Punkt, ...)

- Základní údaje
- Obsah korpusu
- Jádro
- Filmové titulky

Obsah (verze 13)

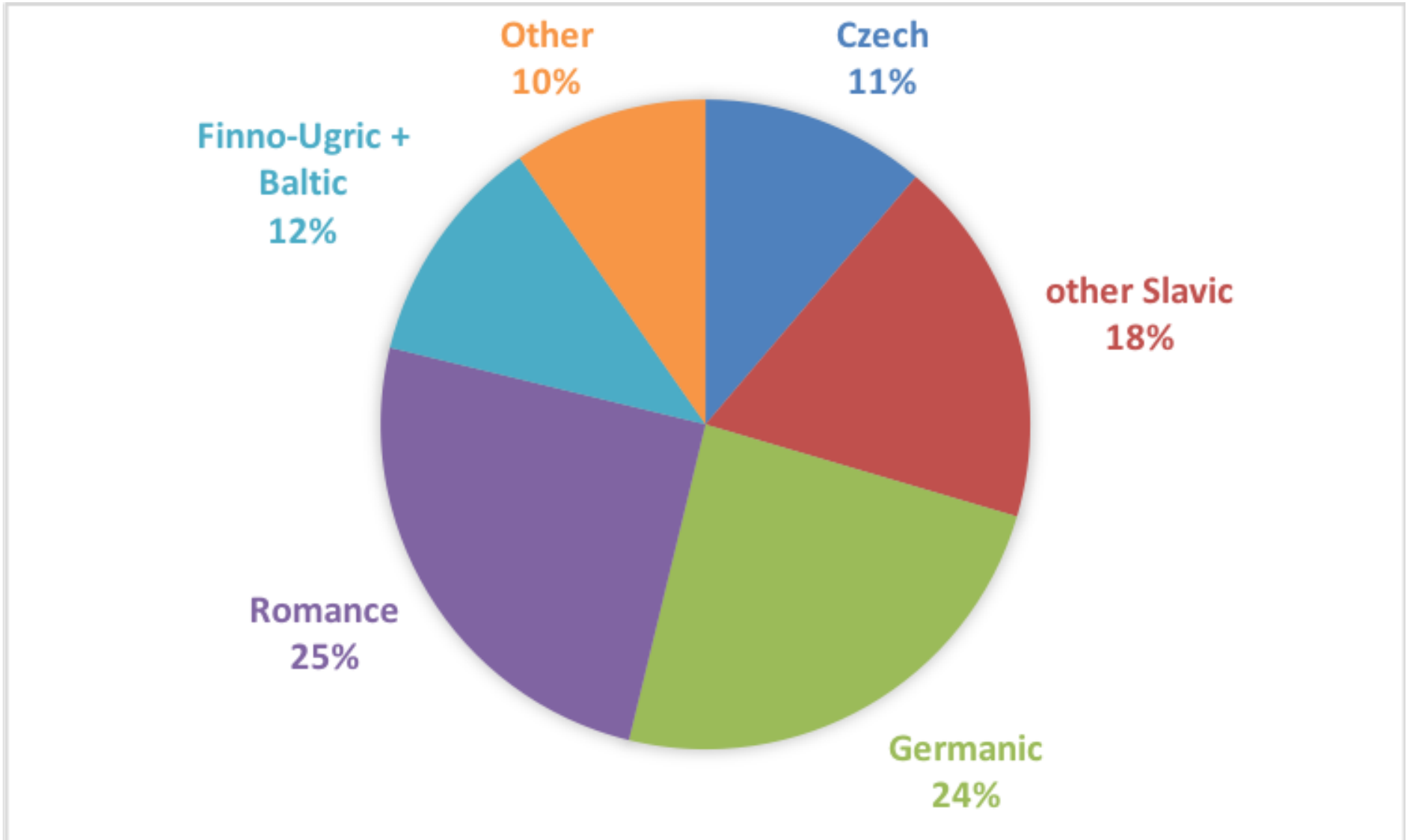
40 jazyků + čeština

- 10 slovanských: be, bg, hr, mk, pl, ru, sk, sl, sr, uk
- 7 germánských: da, de, en, is, nl, no, sv
- 6 románských: ca, es, fr, it, pt, ro
- 5 ugrofinských + baltských: et, fi, hu, lt, lv
- 12 ostatních: ar, el, he, hi, ja, ms, mt, rn, sq, tr, vi, zh

☛ Jen málo textů je k mání ve více než 20 jazycích

☛ Jazyky se velmi liší objemem textů

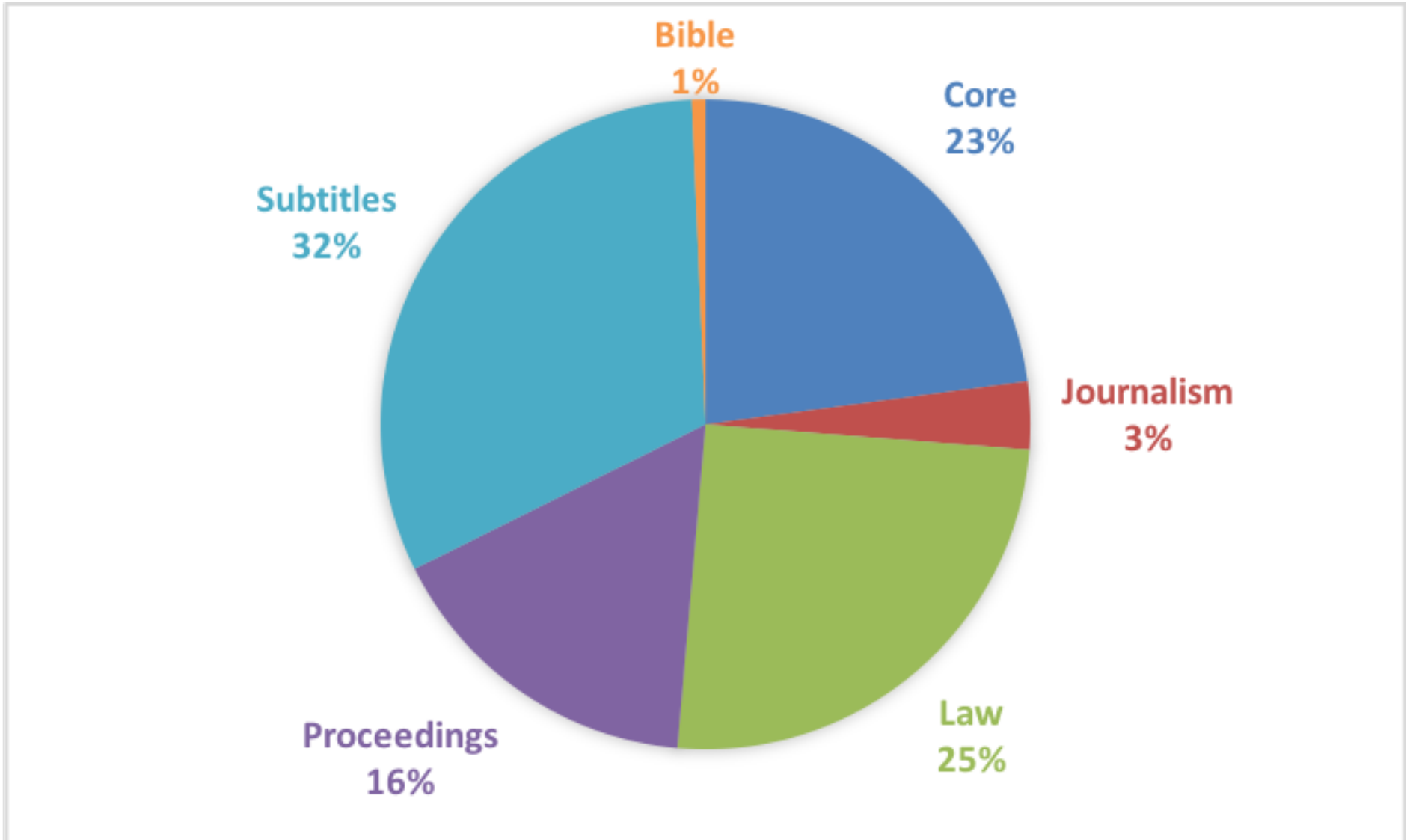
Skupiny jazyků



Druhy textů

- **Celkem** – skoro 1.8 miliardy slov
- **Beletrie** – také literatura faktu, zkorigováno, tzv. **jádro**
- **Kolekce** – volně dostupné texty
  - **Žurnalistika**
    Project Syndicate http://www.project-syndicate.org/
    VoxEurope http://www.voxeurop.eu/
  - **Právo**
    Acquis Communautaire
    http://langtech.jrc.ec.europa.eu/JRC-Acquis.html
  - **Jednání parlamentu**
    Europarl http://www.statmt.org/europarl/
  - **Filmové titulky**
    Open Subtitles http://www.opensubtitles.org
  - **Bible**

Druhy textů


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

Vývoj

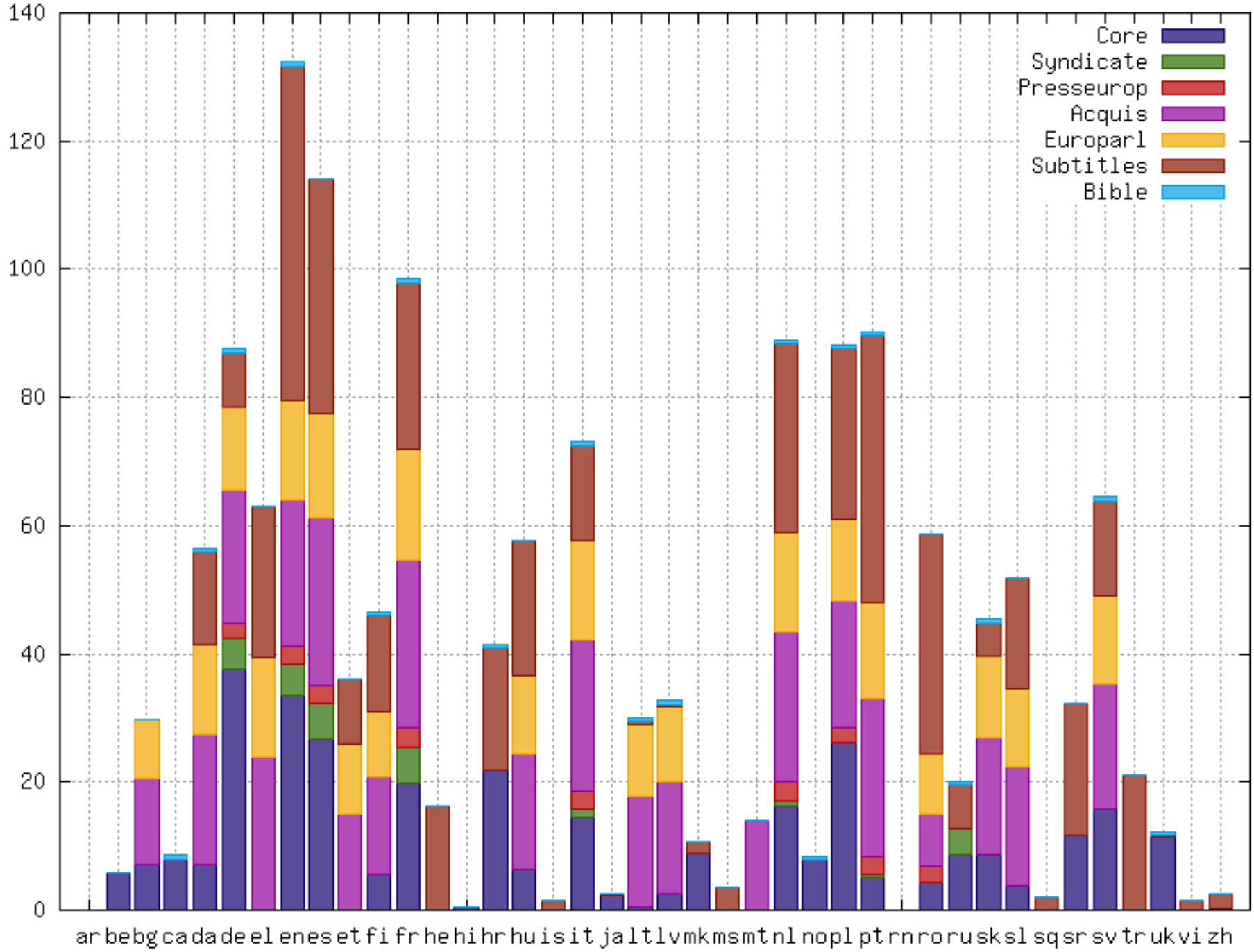
- Údaje pro všechny jazyky kromě češtiny; v.13: 204 M slov

v.	rok	M slov	jazyky	značky	novinky
0	2008	25	19	0	ParaConc, Park
1	2009	35	20	10	značky, lemmata
2	2009	49	21	10	*Project Syndicate*, jednojazykové korpusy
3	2011	72	22	13	oddělené zarovnání
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	plánování textů
10	2017	1 484	39	23	*The Bible*, Treq v.2
11	2018	1 508	39	26	značky pro ja; pro uk, be UD[1]
12	2019	1 533	40	27	zh, včetně značek
13	2020	1 551	40	27	zh i v jádru + verze anotovaná podle UD

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

Objem textů v milionech slov

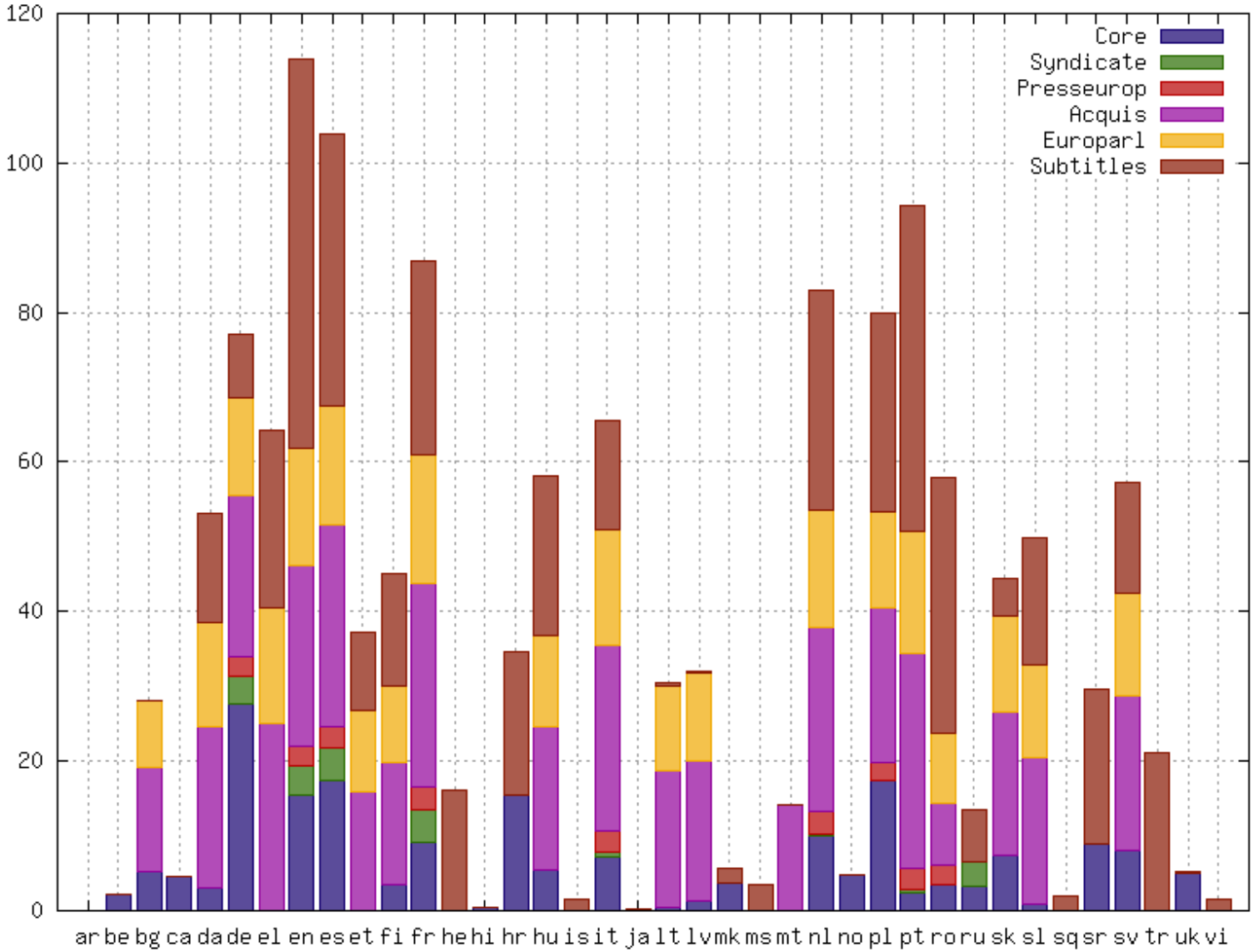
	česky	jinak	**celkem**
jádro	113,8	327,9	**441,7**
žurnalistika	6,7	52,3	**58,9**
právo	19,0	406,5	**425,5**
parlament	12,9	263,9	**276,8**
titulky	50,6	489,2	**539,8**
Bible	0.6	11.5	**12,1**
celkem	203,6	1551,2	**1754,8**


Core
Syndicate
Presseurop
Acquis
Europarl
Subtitles
Bible
0
20
40
60
80
100
120
140
ar be bg ca da de el en es et fi fr he hi hr hu is it ja lt lv mk ms mt nl no pl pt rn ro ru sk sl sq sr sv tr uk vi zh

Počet textů (jen jádro a titulky)

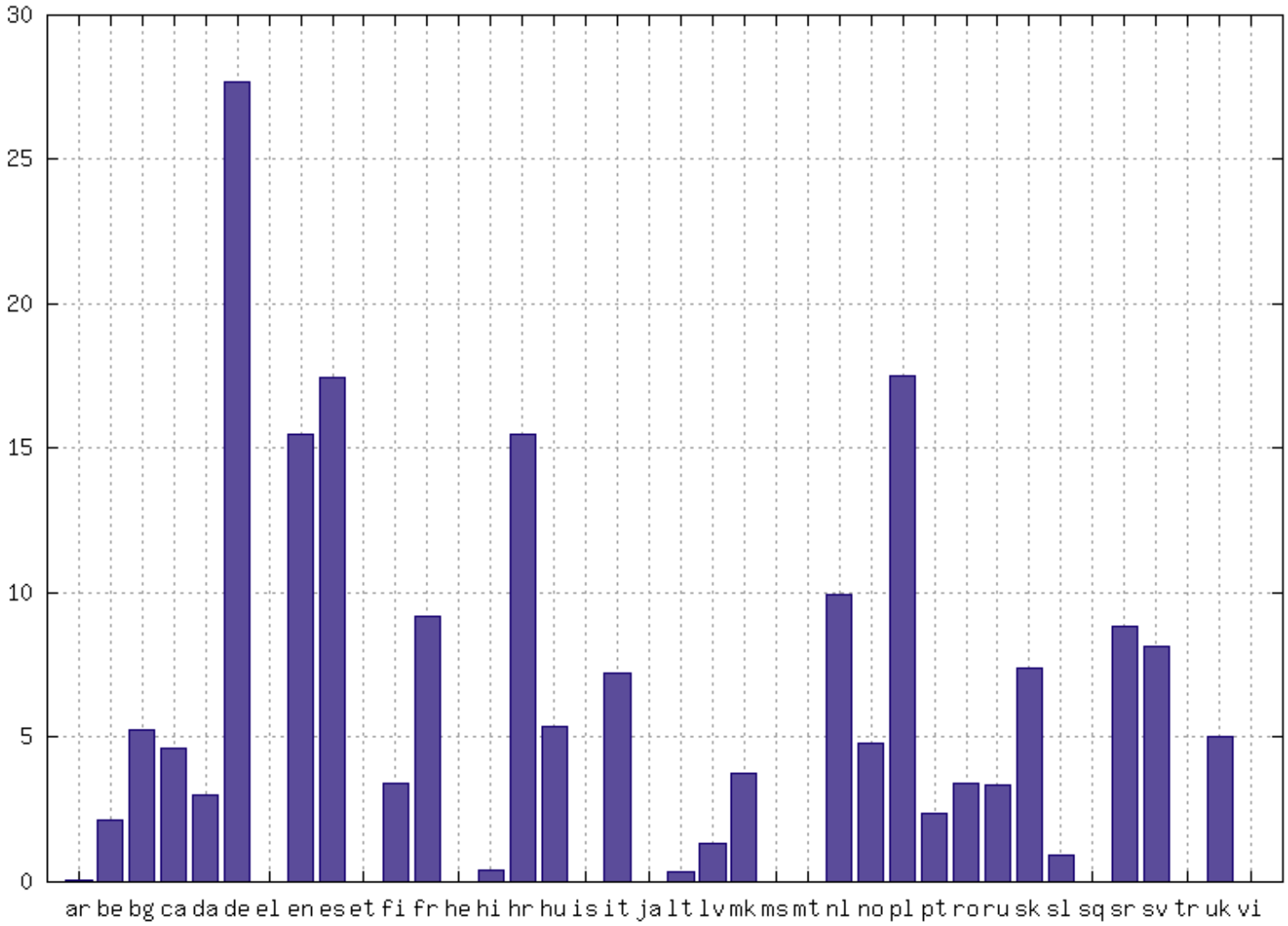
	česky	jinak	**celkem**
jádro	1 656	3 993	**5 649**
titulky	10 400	88 861	**99 261**

Obsah korpusu podle jazyků a typu textů



- Základní údaje
- Obsah korpusu
- Jádro
- Filmové titulky

Jádro (beletrie)

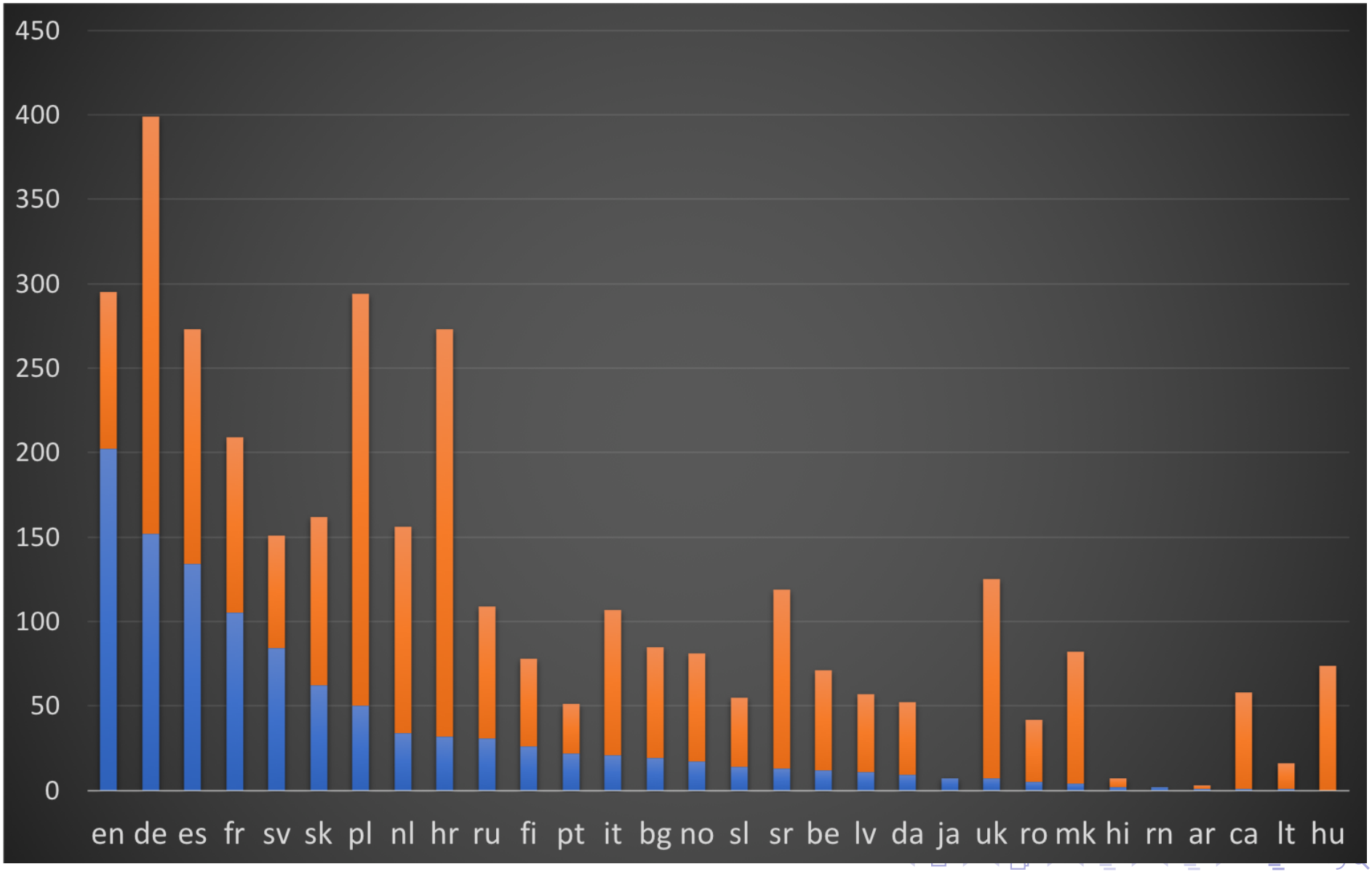


Jádro

Pozor, ne všecky údaje jsou 100% aktuální!

- Průměrný počet jazykových verzí jednoho textu: 3.2
- Za všechny jazyky: 1400 originálů (38%), 3657 překladů
- Za češtinu: 330 originálů (25%), 1327 překladů
- Textů bez originálu: 173

Originály a překlady v jádru



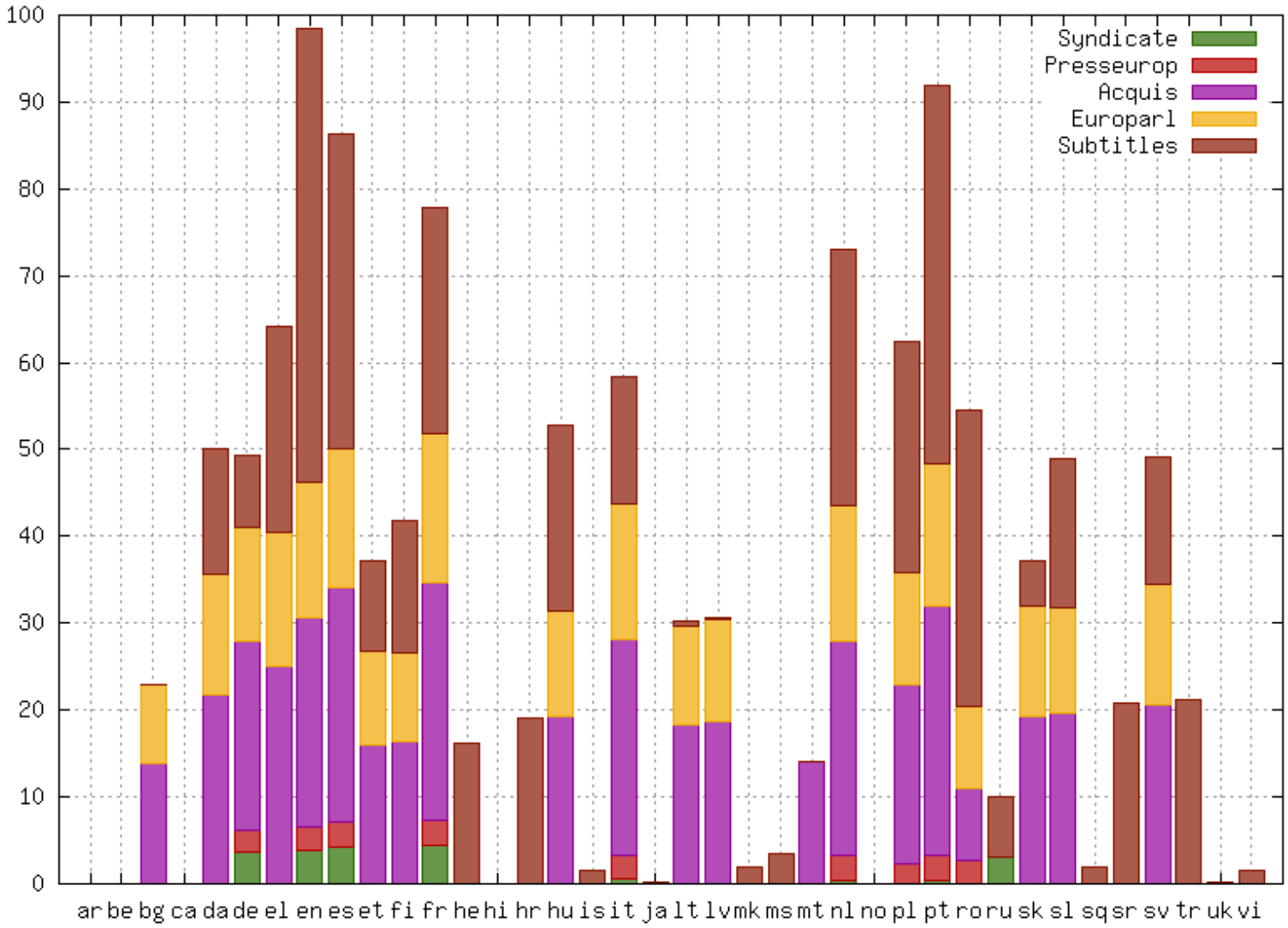
slovanské	jiné	autor	název
11	15	Rowling	*Harry Potter and the Philosopher's Stone*
11	15	Saint-Exupéry	*The Little Prince*
11	12	Carroll	*Alice in Wonderland*
11	12	Orwell	*1984*
11	12	Tolkien	*The Hobbit or There and Back Again*
11	8	Bulgakov	*The Master and Margarita*
11	7	Milne	*Winnie the Pooh*
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*

slovanské	jiné	autor	název
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*
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*
...	...	...	...

Jak vylepšovat jádro?

- **Reprezentativnější/vyváženější**
  typy textů, období, originály/překlady, autoři, překladatelé
  – pro kontrastivní i translatologický výzkum
- **Čím víc, tím líp**
  – průnik textů může být příliš malý třeba i pro němčinu a angličtinu
- **Originály** by měly být vždycky
- **Víc překladů** jednoho textu v jednom jazyku
- **Syntaktická** anotace

Kolekce (žurnalistika, právnické texty, titulky, ...)



- Základní údaje
- Obsah korpusu
- Jádro
- Filmové titulky

Titulky v *InterCorpu*

- Z databáze Open Subtitles https://www.opensubtitles.org
- Filmy i seriály (každý díl zvlášť)
- Jen filmy, ke kterým existují i české titulky
- Je-li víc verzí, vybírá se ta nejlepší (heuristicky)
- Metadata podle kódu IMDb

doc.id	_SUBTITLES	**text.id**	cs:_SUBTITLES:10624_1of1
text.author		**text.title**	Rasuto ran: Ai to uragiri no hyaku-oku en - shissô Feraari 250 GTO
text.lang	cs	**text.version**	00
text.group	Subtitles	**text.publisher**	OpenSubtitles
text.pubplace		**text.pubDateYear**	
text.pubDateMonth		**text.origyear**	1991
text.isbn		**text.txtype**	subtitles
text.comment	ID4361	**text.original**	
text.srclang	ja	**text.translator**	
text.transsex		**text.authsex**	
text.transcomment		**text.collectionauthor**	
text.collectiontitle		**text.volume**	
text.pages		**text.lang_var**	
text.wordcount	8786	**p.id**	cs:_SUBTITLES:10624_1of1:1
s.id	cs:_SUBTITLES:10624_1of1:1:530	**hi.rend**	
lb.id		**div.id**	
div.type			

Počet otitulkovaných filmů podle jazyků

cs	10 400	hu	4 461	ro	5 638
da	2 322	is	246	ru	1 329
de	1 453	it	2 554	sk	991
el	4 430	ja	68	sl	3 305
en	7 963	lt	114	sq	318
es	6 604	lv	52	sr	3 870
et	1 931	mk	326	sv	2 598
fi	3 585	ms	603	tr	4 737
fr	4 213	nl	4 899	uk	51
he	3 166	pl	5 496	vi	176
hr	3 676	pt	7 376	zh	310

Proč jsou titulky v korpusu?

- Paralelní mluvené texty není snadné získat, a když, tak jen specifické žánry
- Přímá řeč v beletrii není autentický mluvený jazyk, ale stylizace
- Tlumočené záznamy jednání vznikají ve formální komunikační situaci
- Představě mluveného paralelního korpusu stojí filmové titulky nejblíže
- “kvazi-mluvený” korpus

Specifika amatérských titulků

- Obvykle usilují o věrnost, méně často se v nich vynechává a kondenzuje
- profesionální titulky nesledují vždy přesně zvukovou stopu (dialogovou listinu nebo scénář), vynechávky a kondenzace mohou měnit syntax i lexikální podobu replik
- Často z odposlechu, bez psané předlohy
- Většinou bez jazykové korektury
- Zarovnání se nekontroluje
- Překlad často z jiného jazyka než jazyka originálu
- Mohou být skvělé i hrůzostrašné:
  Here we go. → *Pojď za strejdou sem.*

Využití titulků

- Hledání stylově adekvátních ekvivalentů, zejména víceslovných výrazů [Charciarek(2019)]
- Hovorová češtiny jako cizí jazyk, výuka metodou DDL (*data-driven learning*) [Zasina(2020), Johns(1991)]
- Pro lexikální ekvivalenty se často hodí *treq*

Problémy

- Nedostatečný kontext (příliš krátké věty):
  Hey, Mama. → *Ahoj mamko.*
- Nevyjádřený větný člen nebo zájmeno:
  Mamka ji používá. → *She, uh. . . She usin' it.*
- Kvalita překladu
- Chyby v zarovnání
- Chybí obraz!

Kde si o tom něco přečíst

- Kurs práce s korpusem v sedmi lekcích
  https://wiki.korpus.cz/doku.php/kurz:uvod
- Manuál rozhraní KonText
  https://wiki.korpus.cz/doku.php/manualy:kontext:index
- Hledání v paralelním korpusu
  https://wiki.korpus.cz/doku.php/kurz:hledani_v_paralelnim_korpusu
- Dokumentace ke korpusu InterCorp verze 12
  https://wiki.korpus.cz/doku.php/cnk:intercorp:verze12

- Dotazy on-line
- Využití InterCorpu pro aplikace a výzkum

Dotazy on-line

KonText

- Používá se taky v Lindat/Clarin: https://lindat.mff.cuni.cz/services/kontext/
- Výběr textů:
  - jazyky, verze korpusu
  - název textu, rok vydání, typ textu
  - originál/překlad, jazyk originálu
  - autor, překladatel, ...
- Paralelní dotazy, dotazovací jazyk CQL
- Pozitivní a negativní filtry na konkordance
- Třídění, frekvenční distribuce, kolokace
- Vlastní subkorpusy, export konkordancí

Statistika přístupů

Počet dotazů (2017–2018)

- 553 denně
- 89% bez specifikace typu textů
- 2.6% volí 3 nebo více jazyků, 13.4% jazyk jediný
- Nejčastější dvojice jazyků: en/de↔cs (58%)

Hledání lexikálních ekvivalentů

Treq – databáze překladových ekvivalentů – http://treq.korpus.cz

- Částečně nahrazuje zarovnání po slovech, které v konkordancích chybí
- Dvojice lexikálních ekvivalentů z paralelních textů r zarovnaných po slovech
- cs/en $\leftrightarrow$ libovolný jazyk
- Filtrování podle typu textů
- Hledání forem nebo lemmat
- Podpora pro regulární výrazy
- Jednotlivá slova nebo víceslovné výrazy

Co by ještě bylo třeba

- **biKWiC** – zvýraznění ekvivalentu klíčového slova e
- Informace o **zarovnání**: 1:1 / 2:1 / 1:2 / automatické / manuální / míra spolehlivosti
- Lexikální profily (podobně jako **Word Sketches** [Kilgarriff et al.(2014)])
- **Zarovnání** po slovech, víceslovných jednotkách, větných členech
- **Crowdsourcing** k opravě chyb v textech a anotaci

- Dotazy on-line
- Využití InterCorpu pro aplikace a výzkum

Poskytování úplných textů

- Zachování autorských práv
- Technická ochrana před zneužitím:
  náhodné pořadí bloků překladových dvojic vět
- Bloky dvojic vět o délce max. 100 slov
- Licence pro školství a výzkum, bez možnosti předávání dalším uživatelům

Publikace využívající InterCorp

- https://www.korpus.cz/biblio: 193 položek
- https://ukaz.cuni.cz/: 331 položek
- https://www.researchgate.net/: 87 entries
- https://scholar.google.com: 2 740 položek

Příprava textů

1. Akvizice
2. Skenování a rozpoznávání znaků (OCR)
3. Korektury
4. Segmentace (rozpoznání hranic vět)
5. Zarovnání
6. Kontrola segmentace a zarovnání
7. Morfosyntaktické značkování

Nástroje používané při zpracování textů

1. Bibliografická databáze
   https://intercorp.korpus.cz/DocDatabase/
2. *Intertext* – editor paralelních textů
   https://intercorp.korpus.cz/intertext/
3. *Punkt* – větný segmentátor
   https://www.nltk.org/_modules/nltk/tokenize/punkt.html
4. *Hunalign* – zarovnávač
   https://github.com/danielvarga/hunalign
5. Taggery pro jednotlivé jazyky
   https://wiki.korpus.cz/doku.php/cnk:intercorp:verze13
6. *UDPipe*
   http://ufal.mff.cuni.cz/udpipe/2 [Straka(2018)]

Bibliografická databáze

- Evidence všech titulů – rozpracovaných i hotových
- Odkazy na dostupné české texty, připravené k zarovnání
- Sleduje postup každého textu všemi fázemi přípravy
- Data z databáze se přidávají jako metadata k textům a využívají v KonTextu

InterText

- Editor paralelních textů k opravám:
  - zarovnání po větách
  - struktury textu (segmentace na věty)
  - překlepů apod.
- Obsahuje automatický zarovnávač (*hunalign*)
- Změny ve struktuře českého textu se promítají do všech zarovnání
- Protokolování změn, export, hledání, záložky
- Dvě verze: serverová a lokální
- Podpora pro třídy uživatelů s odlišnými pravomocemi
- Licence GNU GPL v3: http://wanthalf.saga.cz/intertext

non-1:1 | Mode: auto roll (-2) | 10 p/p

151	Když to považím, strýc císaře pána, a voni ho zastřelejí.		Wenn ich mir das so überleg, ein Onkel Seiner Majestät des Kaisers, und sie erschießen ihn!
152	Vždyť je to ostuda, jsou toho plný noviny.		Das is ja ein Schkandal, die ganzen Zeitungen sind voll damit.
153	U nás před léty v Budějovicích probodli na trhu v nějaké takové malé hádce jednoho obchodníka s dobytkem, nějakého Břetislava Ludvíka. Ten měl syna Bohuslava, a kam přišel prodávat prasata, nikdo od něho nic nekoupil a každý říkal: 'To je syn toho probodnutýho, to bude asi také pěknej lump.'		Bei uns in Budweis hat man vor Jahren auf dem Markt bei irgendeinem kleinen Streit einen Viehhändler erstochen, einen gewissen Bretislav Ludwig, der hatte einen Sohn namens Bohuslav, und wenn der seine Schweine verkaufen kam, wollt niemand was von ihm kaufen, und jeder hat gesagt: ‚Das ist der Sohn von diesem Erstochenen. Das wird gewiß auch ein feiner Lump sein.
154	Musel skočit v Krumlově z toho mostu do Vltavy a museli ho vytáhnout, museli ho křísit, museli z něho pumpovat vodu a von jim musel skonat v náručí lékaře, když mu dal nějakou injekci."		Er hat in Krummau von der Brücke in die Moldau springen müssen, und man hat ihn wieder zu Bewußtsein bringen müssen, und man hat aus ihm das Wasser herauspumpen müssen, und er hat in den Armen des Arztes seinen Geist aufgeben müssen, wie der ihm irgendeine Injektion gemacht hat.“
155	"Vy ale máte divná přirovnání," řekl Bretschneider významně, "mluvíte napřed o Ferdinandovi a potom o obchodníku s dobytkem."		„Sie ziehen aber merkwürdige Vergleiche“, sagte Bretschneider bedeutungsvoll, „zuerst sprechen Sie von Ferdinand und dann von einem Viehhändler.“
156	"Ale nemám," hájil se Švejk, "bůh mě chraň, abych já chtěl někoho k někomu přirovnávat.		„I wo“, verteidigte sich Schwejk. „Gott bewahre, daß ich jemand mit jemandem vergleichen möcht.
157	Pan hostinský mne zná.		Der Herr Wirt kennt mich.
158	Viď že jsem nikdy nikoho k někomu nepřirovnával?		Nicht wahr, ich hab nie jemanden mit jemandem verglichen?
159	Já bych jenom nechtěl být v kůži té vdovy po arcivévodovi.		Ich möcht nur nicht in der Haut der Frau Erzherzogin stecken.
160	Co teď bude dělat?		Was wird die jetzt machen?

Lingvistická anotace

- Tokeny
- Lemmata
- Morfosyntaktické značky
- *Syntaktické funkce a struktura*

Současné řešení – používáme dostupné nástroje (taggery), včetně

- Tokenizace spojené s daným nástrojem
- Sady značek navržené pro daný jazyk

Nástroje pro tokenizaci, lemmatizaci a značkování

jazyk	nástroj	předložka determinátor adjektivum substantivum
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

Problémy s různými sadami značek

Hyperonymie / hyponymie

Značka je obecnější než její obdoba v druhém jazyce

- **IN** se v angličtině používá pro
  - předložky i
  - podřadicí spojky,
- ale v ostatních jazycích jsou pro ně dvě značky.

Částečně se překrývající význam

- Odpovídající značky ze dvou znakových sad se shodují jen částečně

Částečný překryv – cs:PD × pl:adj

cs	v	těch	nejodlehlejších	zástavbách
	RR - - 6	**PDXP6**	**AAFP6 - - - - 3A**	**NNFP6 - - - - - A**
pl	w	tym	wspaniałym	apartamencie
	prep:loc:nwok	**adj:sg:loc:m3:pos**	**adj:sg:loc:m3:pos**	**subst:sg:loc:m3**

- české ***těch*** se značkuje jako ukazovací zájmeno, přičemž se nerozlišuje, zda je užito v pozici substantivní nebo adjektivní
- polské ***tym*** se značkuje jako slovo s adjektivním skloňováním

Problémy se značkováním ve více jazycích – souhrn

Značky – rozdíl nejen v kódování

- en: *under*, *because* – **IN** (předložka i podřadicí spojka)
- cs: *těch* – **PD** × pl: *tych* – **adj**
- cs: *devátá* – **Cr** × pl: *dziewiąta* – **adj**
- en: *remotest* – **JJS** × de: *abgelegenste* – **ADJA**

Tokenizace – podobné jevy se řeší jinak

- *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*

Řešení?

- Konverze národních značek v textech do jednotné sady značek
  - Pro některé sady značek konverze není specifikovaná
  - Nekompatibilní tokenizace
  - Ztrátová nebo chybná konverze
- Přeznačkování nástrojem natrénovaným na jednotné sadě značek
  - Může být více chyb kvůli menším trénovacím datům
  - Národní značky v textech nelze zachovat kvůli tokenizaci

Existující standard

Universal Dependencies – http://universaldependencies.org/

- Vytvořený hlavně pro syntax
- Faktický standard také pro morfologické kategorie
- Tokenizace na dvou úrovních: na ortografická a syntaktická slova

Problémy?

- Substantiva jako NOUN nebo PROPN
- Příčestí jako ADJ, NOUN nebo VERB, podle jazyka a kontextu
- Deverbativní substantiva jako VERB nebo NOUN, podle jazyka a kontextu
- Modální slovesa jako VERB nebo AUX, podle jazyka
- Řadové číslovky jako ADJ nebo ADV
- DET pro všechny kvantifikátory a zájmena v prenominální pozici

Některé problémy

- Nelze pracovat s více verzemi jednoho textu v jednom jazyce
  - Technicky se korpus skládá z podkorpusů pro každý jazyk
  - Není jasné, jak by mělo vypadat zadávání dotazů a zobrazování výsledků
- Velké rozdíly mezi jednotlivými jazyky: velikost, značkování, typy textů
- Různá pravidla tokenizace a sady značek pro různé jazyky
- Texty bez české verze
  - Zatím musí mít každý cizí text český protějšek
- Nelze více verzí překladů jednoho textu

Perspektivy

Využití korpusu

- vylepšování vyhledávacího rozhraní
- integrace s jinými paralelními korpusy?

Obsah

- lepší rovnováha mezi jazyky a typy textů
- více jazyků

Anotace

- zlepšování kvality zarovnání a dělení na věty, také pomocí crowdsourcingu (motivace uživatelů k upozorňování na chyby)
- zarovnání po slovech, víceslovných výrazech, větných členech
- zkvalitňování lingvistické anotace:
  - co nejlepší nástroje pro co nejvíce jazyků
  - jednotné zásady tokenizace spřežek a víceslovných výrazů
  - harmonizace značkových sad

Syntaktická anotace

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.


Bojar, O. (2012).
Čeština a strojový překlad (Czech Language and Machine Translation), volume 11 of *Studies in Computational and Theoretical Linguistics*.
Ústav formální a aplikované lingvistiky MFF UK, Praha, Czech Republic.


Bojar, O. & Žabokrtský, Z. (2009).
CzEng0.9: Large parallel treebank with rich annotation.
Prague Bulletin of Mathematical Linguistics, **92**.


Charciarek, A. (2019).
Využití paralelního korpusu v translatologii (na základě česko-polského intercorpu).
Bohemistyka, **19**(2), 194–216.


Johns, T. (1991).
Should you be persuaded: Two samples of data-driven learning materials.
In T. Johns and P. King, editors, *Classroom Concordancing. English Language Research Journal*, volume 4, pages 1–16. University of Birmingham.


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.


Koehn, P. (2020).
Neural Machine Translation.
Cambridge University Press.


Straka, M. (2018).
UDPipe 2.0 prototype at CoNLL 2018 UD shared task.
In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.


Zasina, A. (2020).
Parallel corpus in teching conversational skills in Czech as a foreign language.
In prep.