Filmové titulky jako jazyková data

Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Ústav českého národního korpusu
Filozofická fakulta Univerzity Karlovy

Dobrý překlad není vidět?!
Filozofická fakulta Univerzity Karlovy v Praze
6. října 2020

Obsah

Obsah

Český národní korpus – ČNK

- Jazykový korpus je elektronický soubor autentických textů v přirozeném kontextu, který by měl představovat jazyk tak, jak se opravdu používá. (https://wiki.korpus.cz/)
- ČNK vytváří a udržuje Ústav Českého národního korpusu. (https://ucnk.ff.cuni.cz/)
- *1994
- Řada korpusů, dostupných on-line přes jednotné vyhledávací rozhraní a další nástroje
- Psané, mluvené, paralelní, historické, žákovské, webové

Co všechno obsahuje Český národní korpus

- Psaná čeština:
  - SYN2000, SYN2005, SYN2010, SYN2015 – každý 100 mil. slov
  - SYN v.8 – 4,8 miliardy slov
- Mluvená čeština:
  - ORAL – 5,4 mil. slov
  - ORTOFON v.1 – 1 mil. slov
  - BMK, DIALEKT, PMK – nářečí
- **Paralelní korpus**: čeština a 40 dalších jazyků
  - **InterCorp** v.13 – 1.8 miliardy slov
- Webové korpusy – vícejazyčné
  - Aranea – 1 miliarda slov
  - deWac, frWac, itWac, ukWac – 1.3–1.9 miliardy slov
- Historická čeština:
  - DIAKORP v.6 – 3.4 mil. slov
- Další:
  - žákovské, překladatelská čeština, ručně anotované, jiné jazyky

https://wiki.korpus.cz/doku.php/en:cnk:uvod

LOGIN

Několik možností:

- Po (bezplatné) registraci: https://www.korpus.cz/signup
- Přes institucionální login: https://www.korpus.cz/login
- Bez loginu to jde taky, ale možnosti jsou omezené

Jak v ČNK něco najít – konkordance

KonText

http://kontext.korpus.cz

- Vybrat korpus
- Zadat dotaz (základní, lemma, slovní spojení, CQL)
- Poozitivní a negativní filtry konkordancí
- Export konkordancí
- Třídění, frekvenční distribuce, kolokace
- Uživatelské subkorpusy

Obsah

Kdy je korpus paralelní?

- Stejný text ve více verzích (jazyky, verze překladu, ...)
- Zarovnání po větách, odstavcích, někdy i po slovech
- Využití:
  - strojový překlad, vyhledávání v jiném jazyce, promítání anotace
  - podpora překladatelů (CAT)
  - výuka cizích jazyků
  - lexikografie
  - jazykověda
  - jako doplněk nebo náhrada běžného slovníku

Co nabízí paralelní korpus

- Překlad by měl zachovávat význam
- Paralelní kontext
  - → explicitní vyjádření překladové ekvivalence
  - → implicitní anotace významu

- Od významu k výrazu:
  - vyhledat ekvivalentní výraz v jiném nebo stejném jazyce
  - – překladové a kontrastivní studie, výuka cizích jazyků, strojový překlad, podpora překladatelů (CAT)
- Od výrazu k významu:
  - zjistit význam pomocí cizojazyčných ekvivalentů
  - – porozumění textu, promítání anotace, jednojazyková lexikografie

Obsah

Základní údaje

- Část *Českého národního korpusu*
- http://www.korpus.cz/intercorp/
- *2005 (zásluhou Františka Čermáka)
- Zpočátku jako služba jazykovým a filologickým katedrám FF UK
- On-line od roku 2008
- Každý rok nová verze, verze 13 už brzy

Architektura *InterCorp*u

- Zarovnání: po větách
- Každý text je česky a aspoň v jednom dalším jazyce
- Ve většině jazyků má každé slovo morfologickou značku (tag) a základní tvar (lemma)
- Bibliografické údaje (metadata) ke každému textu

Jak v paralelním korpusu něco najít – konkordance

KonText

http://kontext.korpus.cz

- Vybrat verzi korpusu *InterCorp*
- Omezit hledání na určité texty:
  - jazyky, skupina textů (např. Subtitles), rok vydání
  - originál/překlad, jazyk originálu
  - autor
- Paralelní dotazy
- Jinak stejné možnosti jako u jednojazykových korpusů

Lexikální ekvivalenty

Treq – a database of translation equivalents

http://kontext.korpus.cz

- Databáze je vygenerována z paralelních textů zarovnaných po slovech
- cs/en $\leftrightarrow$ libovolný další jazyk
- Omezení na skupinu textů (Subtitles)
- Dotaz pomocí tvaru nebo lemmatu
- Podpora regulárních výrazů
- Dotaz na jednotlivá slova nebo víceslovné výrazy

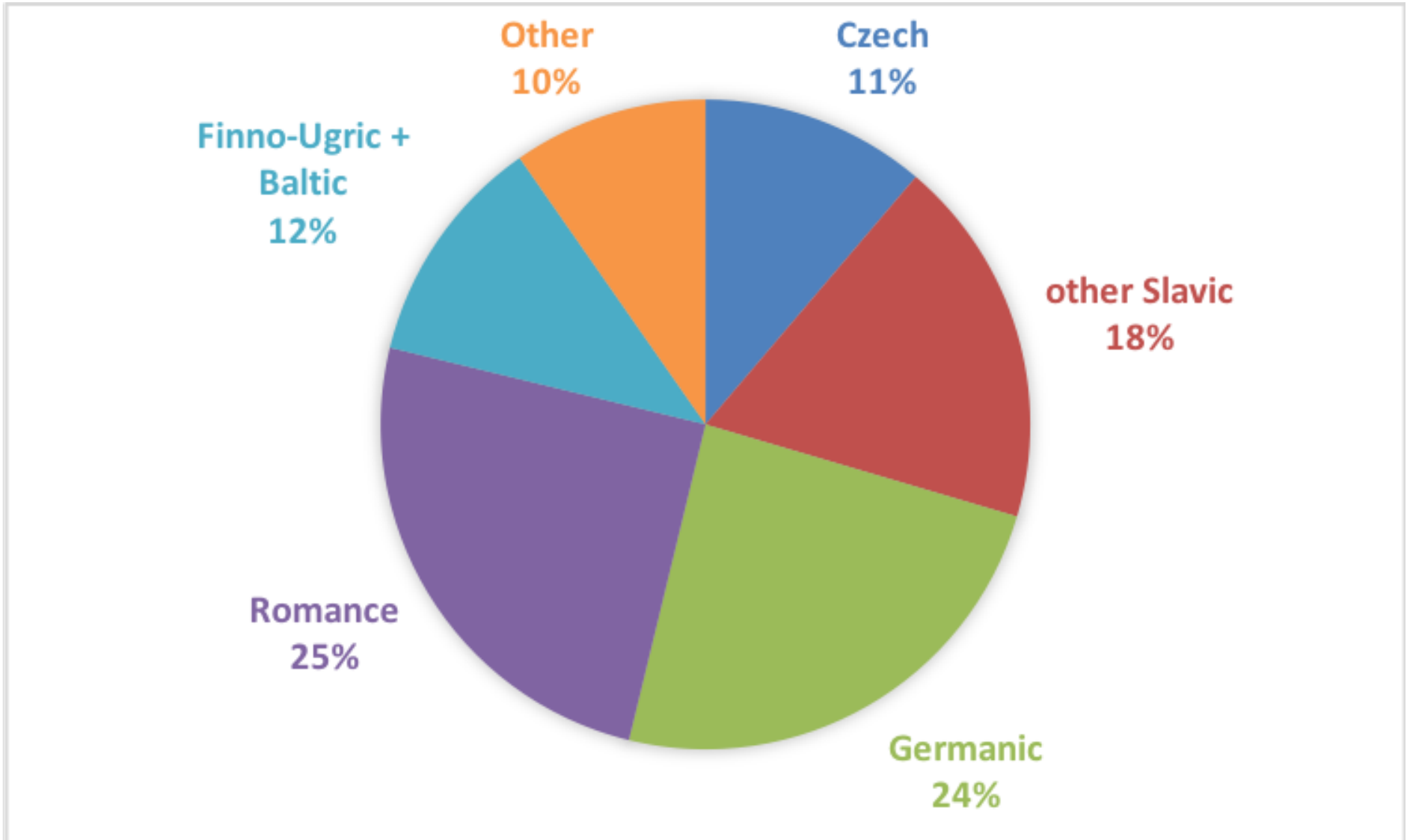
Obsah (verze 13)

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

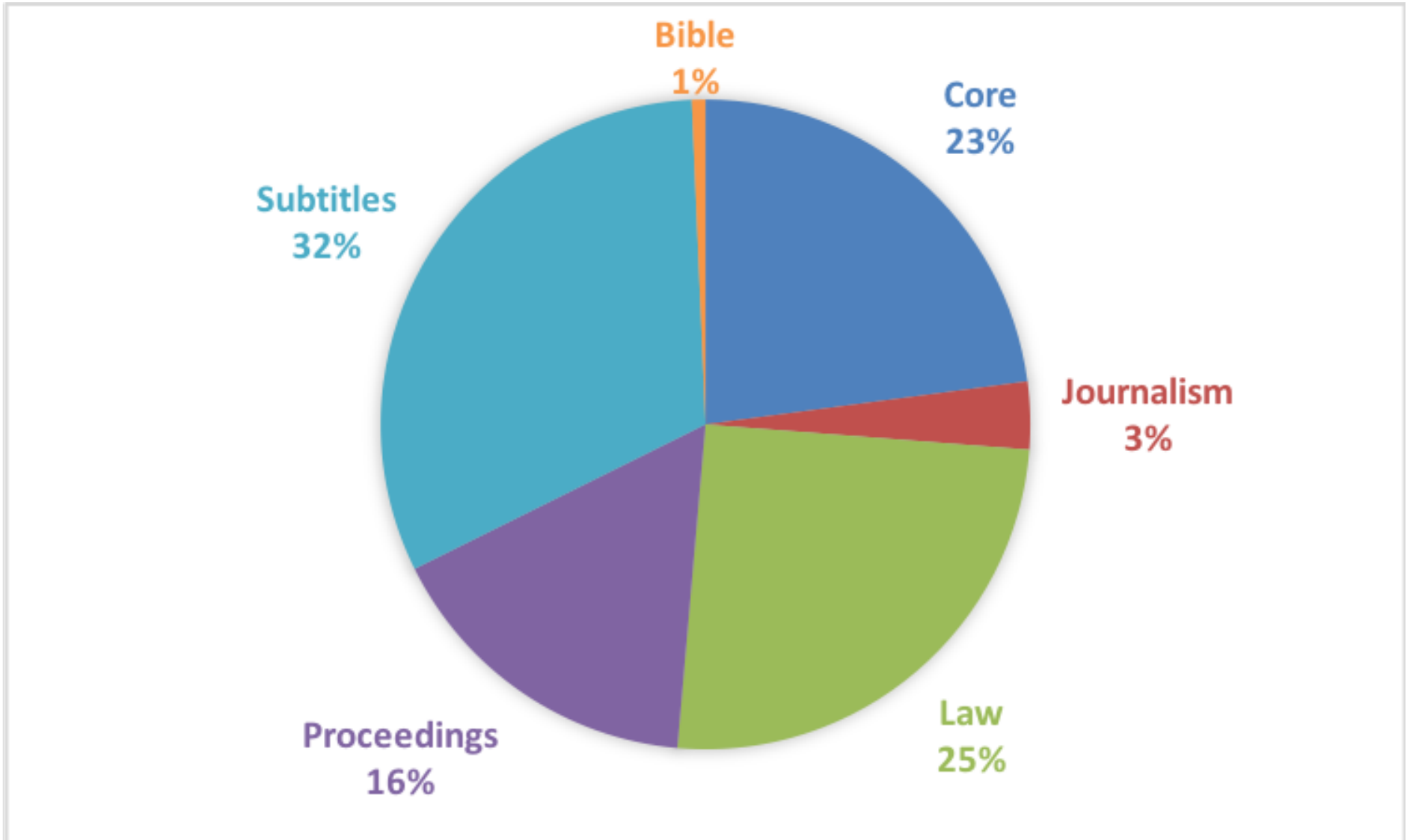
Skupiny jazyků



Druhy textů

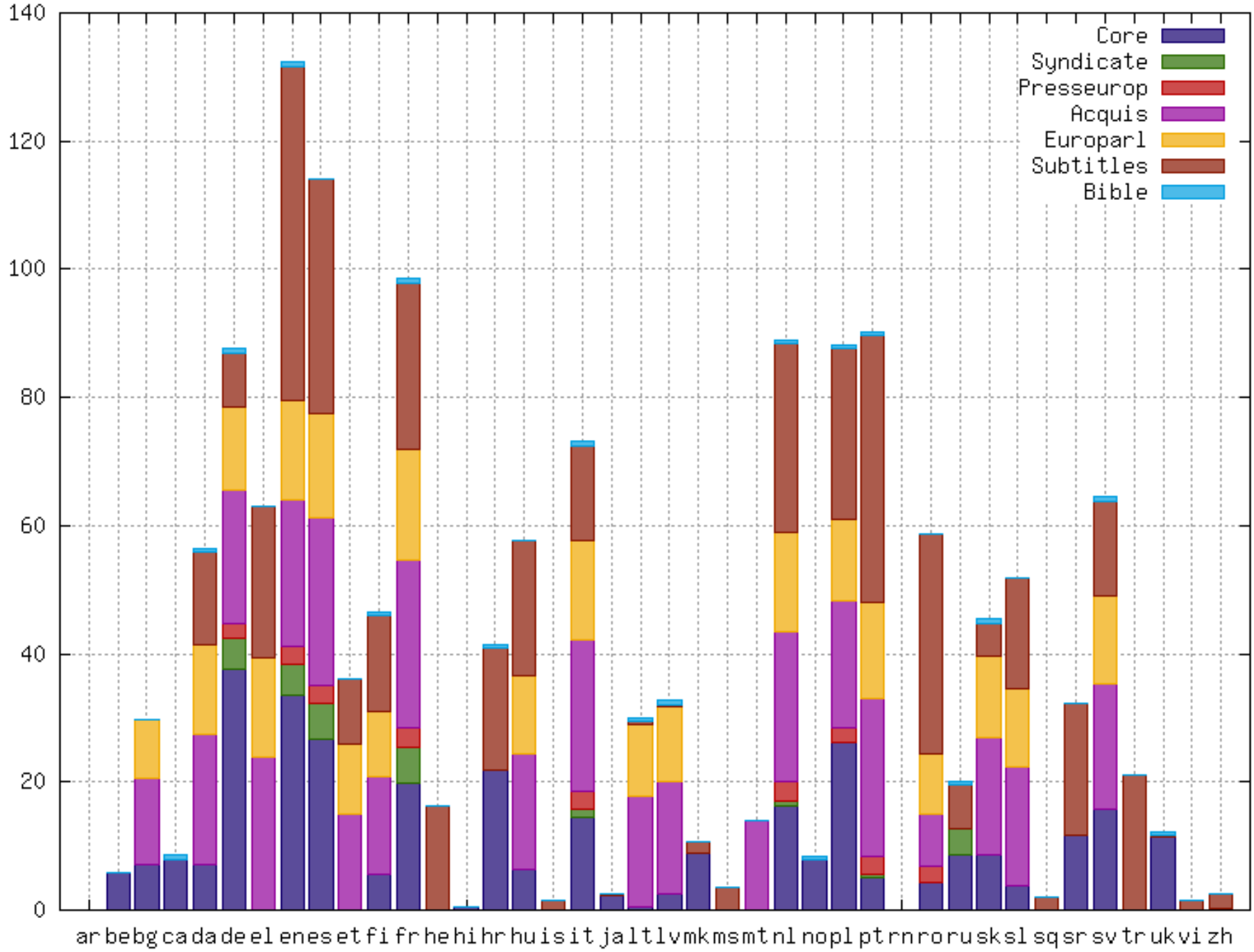
- **Celkem** – skoro 1.8 miliardy slov
- **Beletrie** – také literatura faktu, zkorigováno, tzv. **jádro**
- **Kolekce** – volně dostupné texty
  - **Žurnalistika**
    Project Syndicate http://www.project-syndicate.org/
    VoxEurope http://www.voxeurop.eu/
  - **Právo**
    Acquis Communautaire
    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%

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

Titulky v *InterCorp*u

- 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 Feraar 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:
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 zaznamenávající jazyk v běžných situacích 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 kondenzace mohou měnit syntax i lexikální podobu replik
- Překládají se často z odposlechu, bez psané předlohy
- Často chybí jazyková korektura
- Zarovnání se nekontroluje
- Často se překládají z jiného jazyka než jazyka originálu, aniž je to výslovně uvedeno
- 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. $\rightarrow$ *Ahoj mamko.*
- Nevyjádřený větný člen nebo zájmeno:
  Mamka ji používá. $\rightarrow$ *She, uh. . . She usin' it.*
- Kvalita překladu
- Chyby v zarovnání
- Chybí obraz!


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

Obsah

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, page 1–16. University of Birmingham.

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