Koncepce a zdroje korpusu *InterCorp*

Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Filozofická fakulta Univerzity Karlovy v Praze

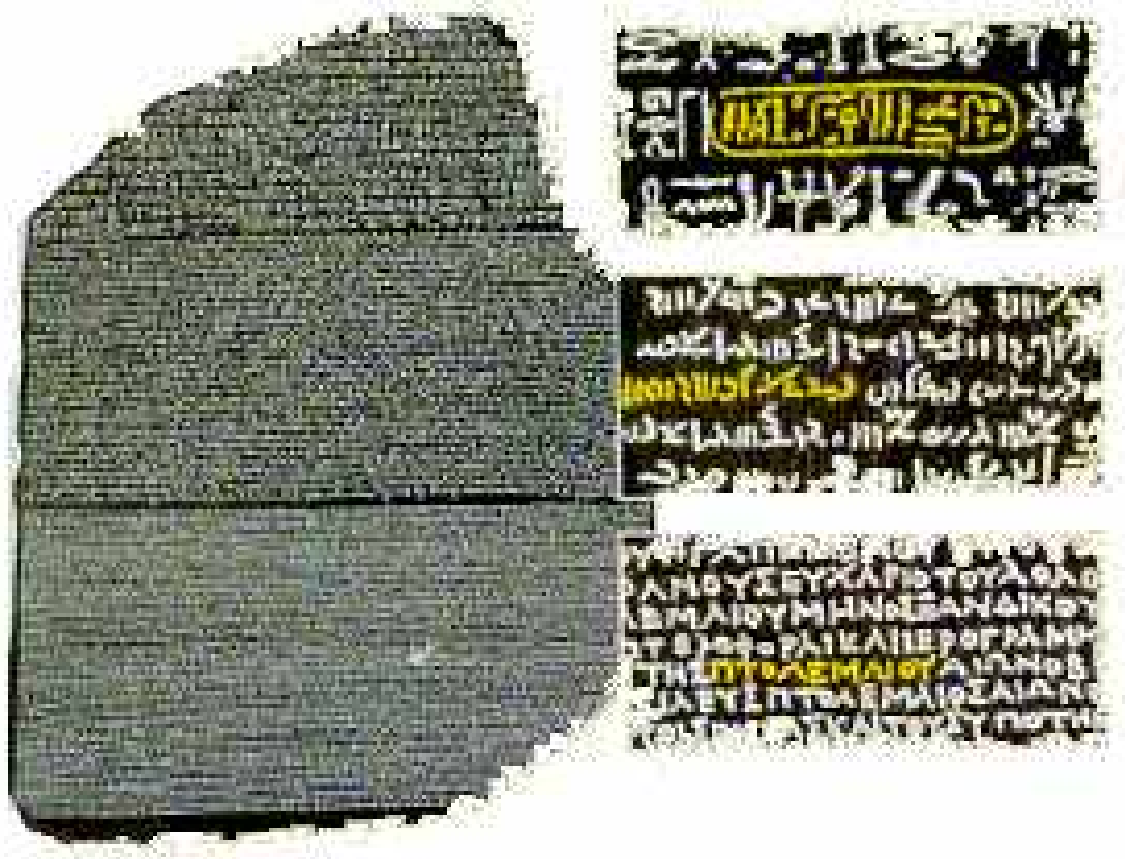
Workshop InterCorp
Praha, 6. září 2013

Osnova

1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Co je to paralelní korpus?

- Paralelní korpus obsahuje stejná nebo srovnatelná data ve více podobách, které se liší jazykem nebo verzí překladu.



Podmínky pro rozumnou práci s paralelními korpusy:

- zarovnání po větách
  - přiřazení překladu věty k jejímu originálu
- paralelní korpusový manažer

Nevýhody paralelních korpusů:

- texty nejsou autentické, většinou jen překlady
- porovnání ekvivalentů může být obtížné, když je překlad hodně volný nebo z třetího jazyka
- korpus není reprezentativní,
  paralelně lze získat jen některé typy textů
- předpokladem rozumného využití je spolehlivé zarovnání po větách, ale:
  - automatické metody zarovnávání nefungují na 100 %
  - zarovnávat ručně je dřina a pro velké texty je to nereálné

Osnova

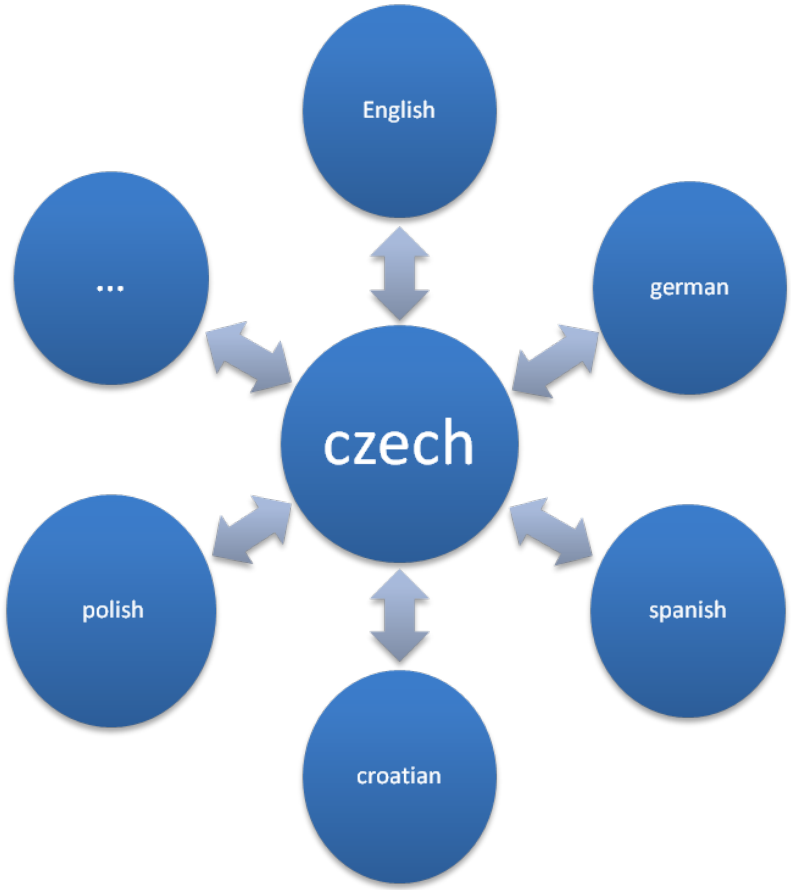
1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Základní údaje

- *InterCorp* – vícejazykový paralelní korpus zaměřený na češtinu
- součást *Českého národního korpusu*
- `http://www.korpus.cz/intercorp/`
- * 2005 jako služba pro lingvistická pracoviště FF UK
- +/– každý rok nové vydání
- už delší dobu se hodně využívá i mimo univerzitní prostředí
- od roku 2012 financován z programu *Velké infrastruktury pro výzkum, experimentální vývoj a inovace*

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 – pokud na to máme nástroje



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

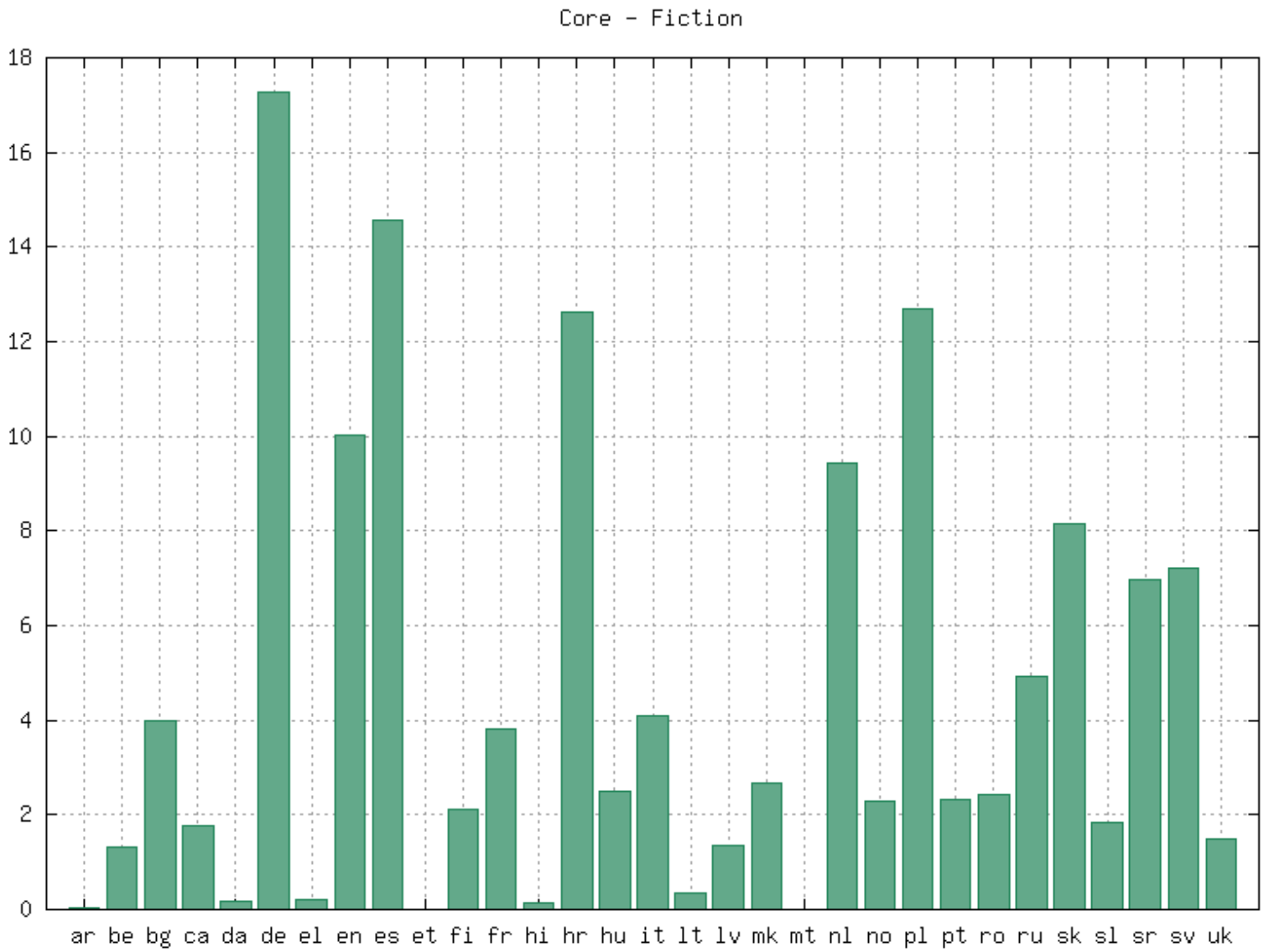
Spolupráce

- 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, ...)

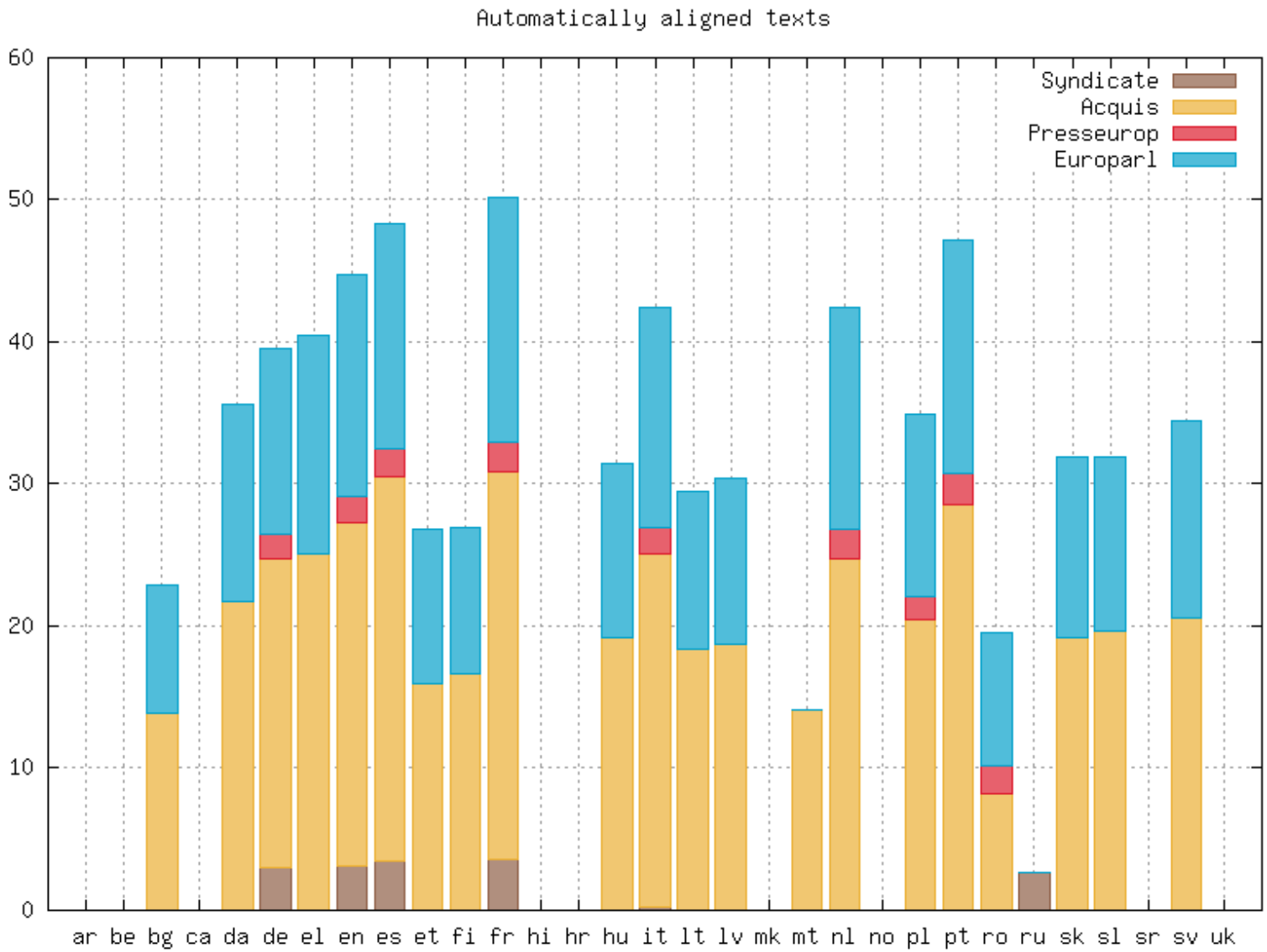
Obsah korpusu – 6. vydání

- **Počet jazyků**: 31 + česky
  - jen málo textů je k mání ve více než 5 jazycích
  - jazyky se velmi liší objemem textů
- **Celková velikost** – 867/100 mil. slov (cizí/české)
- **Jádro** – 139/62 mil. slov: beletrie s manuálně zkorigovaným zarovnáním
- **Kolekce** – texty s automatickým zarovnáním:
  - **Žurnalistika** – 33/4 mil. slov:
    Project Syndicate `http://www.project-syndicate.org/`
    Presseurop `http://www.presseurop.eu/`
  - **Právnické texty** – 430/20 mil. slov:
    Acquis Communautaire
    `http://langtech.jrc.ec.europa.eu/JRC-Acquis.html`
  - **Zápisy z jednání parlamentu** – 265/13 mil. slov:
    Europarl `http://www.statmt.org/europarl/`

Jádro (beletrie)



Kolekce (žurnalistika, právnické texty, ...)



	Jazyk	Jádro	Syndicate	PressEU	Acquis	Europarl	Celkem
ar	arabština	29	0	0	0	0	29
be	běloruština	1 308	0	0	0	0	1 308
bg	bulharština	3 979	0	0	13 816	9 083	26 879
ca	katalánština	1 758	0	0	0	0	1 758
da	dánština	190	0	0	21 680	13 916	35 785
de	němčina	**17 256**	3 050	1 715	21 724	13 089	56 835
el	řečtina	210	0	0	25 070	15 404	40 683
en	angličtina	10 019	3 083	1 863	24 208	15 580	54 753
es	španělština	**14 552**	3 479	1 948	27 001	15 885	62 865
et	estonština	0	0	0	15 963	10 900	26 862
fi	finština	2 131	0	0	16 667	10 241	29 040
fr	francouzština	3 816	3 535	2 054	27 352	17 178	53 936
hi	hindština	155	0	0	0	0	155
hr	chorvatština	**12 625**	0	0	0	0	12 625
hu	maďarština	2 511	0	0	19 168	12 307	33 985
it	italština	4 081	247	1 893	24 850	15 489	46 560
lt	litevština	358	0	0	18 433	11 020	29 811
lv	lotyština	1 337	0	0	18 745	11 689	31 770
mk	makedonština	2 664	0	0	0	0	2 664
mt	maltština	0	0	0	14 133	0	14 133
nl	nizozemština	9 426	0	2 082	24 746	15 563	51 817
no	norština	2 301	0	0	0	0	2 301
pl	polština	**12 710**	0	1 660	20 464	12 805	47 640
pt	portugalština	2 318	0	2 103	28 599	16 481	49 502
ro	rumunština	2 433	0	1 917	8 200	9 446	21 995
ru	ruština	4 937	2 651	0	0	0	7 588
sk	slovenština	8 152	0	0	19 222	12 734	40 108
sl	slovinština	1 855	0	0	19 646	12 241	33 741
sr	srbština	6 972	0	0	0	0	6 972
sv	švédština	7 205	0	0	20 615	13 874	41 694
uk	ukrajinština	1 493	0	0	0	0	1 493
Celkem		138 779	16 044	17 237	430 300	264 926	867 287
cs	čeština	61 962	2 741	1 639	20 285	12 920	99 547

Tituly s nejvyšším počtem verzí I

26	J. K. Rowling *Harry Potter a kámen mudrců*
24	A. de Saint-Exupéry *Malý princ*
21	Lewis Carroll *Alenka v říši divů*
20	Milan Kundera *Nesnesitelná lehkost bytí*
20	J. K. Rowling *Harry Potter a tajemná komnata*
19	Douglas Adams *Stopařův průvodce po galaxii*
19	Milan Kundera *Žert*
18	Dan Brown *Šifra Mistra Leonarda*
18	Michail Bulgakov *Mistr a Markétka*
18	Jaroslav Hašek *Osudy dobrého vojáka Švejka*
18	A. A. Milne *Medvídek Pú*
18	J. K. Rowling *Harry Potter a vězeň z Azkabanu*
18	J. R. R. Tolkien *Pán prstenů I–III*
17	Paolo Coelho *Alchymista*
17	J. R. R. Tolkien *Hobit*

Tituly s nejvyšším počtem verzí II

16	Umberto Eco *Jméno růže*
16	Franz Kafka *Proces*
16	George Orwell *1984*
16	J. K. Rowling *Harry Potter a ohnivý pohár*
15	Anna Franková *Deník*
14	Rudyard Kipling *Kniha džunglí*
14	Milan Kundera *Nesmrtelnost*
14	Nikolaj Ostrovskij *Jak se kalila ocel*
13	Bohumil Hrabal *Příliš hlučná samota*
13	Milan Kundera *Směšné lásky*
13	J. K. Rowling *Harry Potter a Fénixův řád*
12	F. S. Fitzgerald *Velký Gatsby*
12	Bohumil Hrabal *Obsluhoval jsem anglického krále*
12	Franz Kafka *Zámek*
12	Franz Kafka *Proměna*
12	Milan Kundera *Valčík na rozloučenou*

Tituly s nejvyšším počtem verzí III

11	Ernest Hemingway *Stařec a moře*
11	Stanisław Lem *Solaris*
11	Astrid Lindgren *Pipi Dlouhá punčocha*
11	Astrid Lindgren *Karkulín ze střechy*
11	G. G. Marquez *Kronika ohlášené smrti*
11	George Orwell *Farma zvířat*
11	Michal Viewegh *Výchova dívek v Čechách*
10	Karel Čapek *Válka s mloky*
10	Stieg Larsson *Muži, kteří nenávidí ženy*
10	Vladimir Nabokov *Lolita*
10	H. G. Wells *Stroj času*
10	H. G. Wells *Válka světů*
9	Václav Havel *Dálkový výslech*

Osnova

1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

OPUS – an open source parallel corpus

`http://logos.uio.no/opus/`

- 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)*

ASPAC – the Amsterdam Slavic Parallel Corpus

- autor: Adrie Barentsen
- *InterCorp* ho obsahuje téměř celý
- celková velikost >4 mil. tokenů (slov včetně interpunkce)
- 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


Mali Kraljič
Winij Puw
WIE DER STAHL GEHÄRTET WURDE
HARRIUS POTTER
NABOKOV ALICE IN WONDERLAND
Аня в стране чудес
J. R. R. TOLKIEN HOBIT
MEISTER UND MARGARETA
Douglas Adams Stopárov
RUDYARD KIPLING
Dan Brown Da Vinciho kód
Anne Frank dziennik
ROK 1984 GEORGE ORWELL
SOLARIS
Karkulín ze střechy
PIPI DUGA ČARAPA
Knjiga o džungli
Paulo Coelho
ALKIMIST
A. A. MILNE
George Orwell De boerderij der dieren
Edgar Allan Poe Compleet Het volledige proza
Том Сойер
Carlo Collodi PINOCCHIOVA DOBRODRUŽSTVÍ
helen fielding dnevnik bridžet džouns
Jaroslav Hašek DOŽIVLJAJI DOBROG VOJNIKA ŠVEJKA
NABOKOV Lolita
ILF PETROV 12 STOLIČIEK
MIJA
STIEG LARSSON MĚŽI, KTEŘÍ NENÁVIDĚLI ŽENY
Kapetanova kći
DAN BROWN ANDĚLÉ a DÉMONI
UMBERTO ECO DE NAAM VAN DE ROOS
William Golding GOSPODAR MUHA
ATENTAT Harry Mulisch
JEROME K. JEROME TRZECH PANÓW W ŁÓDCE
Patrick Süskind Das Parfum
KLASYKA
Pies Baskerville'ów
ПРОЦЕС
ФРАНЦ КАФКА МЕТАМОРФОЗАТА
The Doll
Lord Jim
ERICH KÄSTNER EMIL I DETEKTIVI
QUO VADIS?
Хенрик Сјенкјевич
Milorad Pavić

ParaSol: A Parallel Corpus of Slavic and other languages

- autoři: Ruprecht von Waldenfels (Bern) a Roland Meyer (Regensburg)
- 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

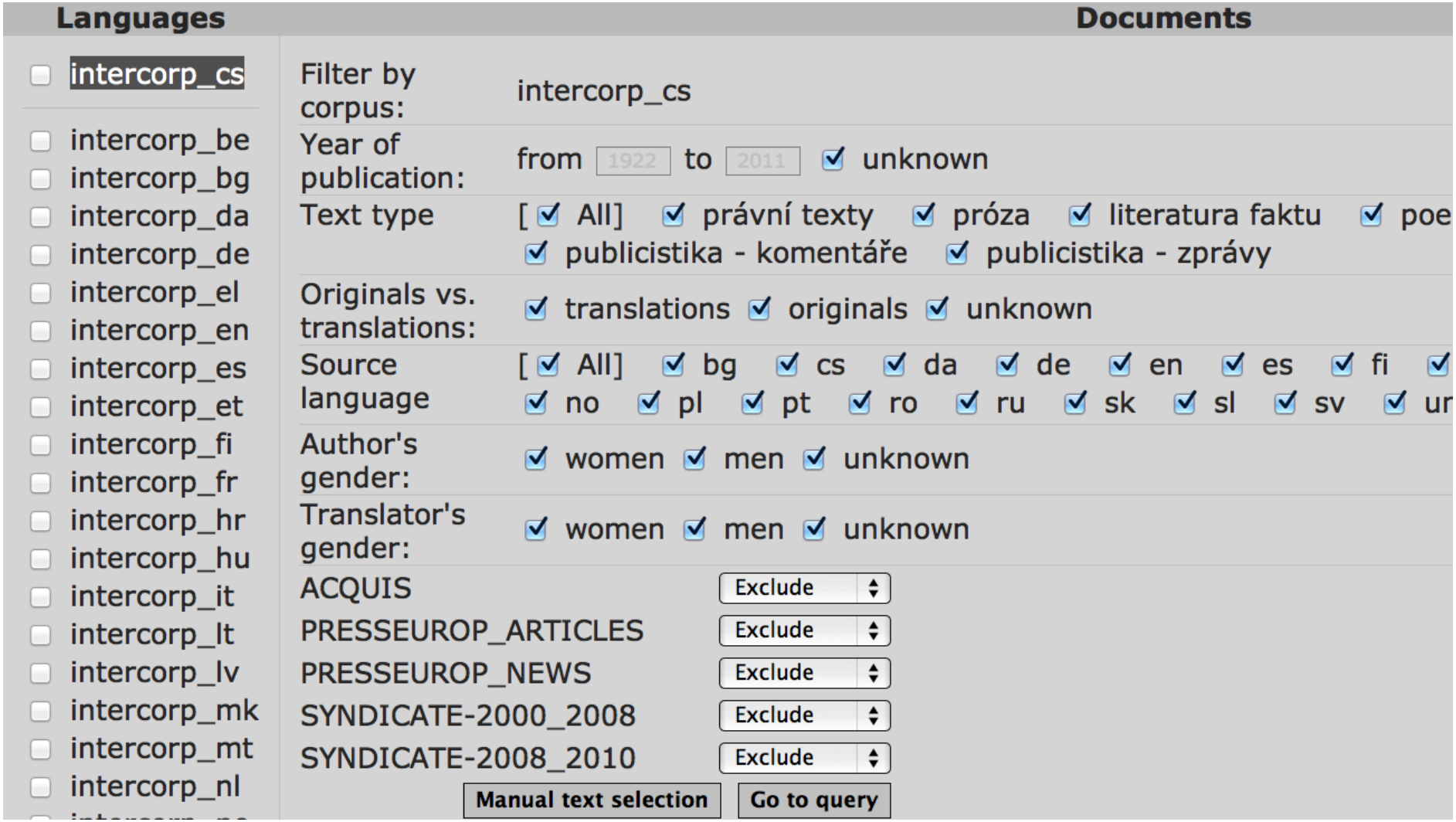
Osnova

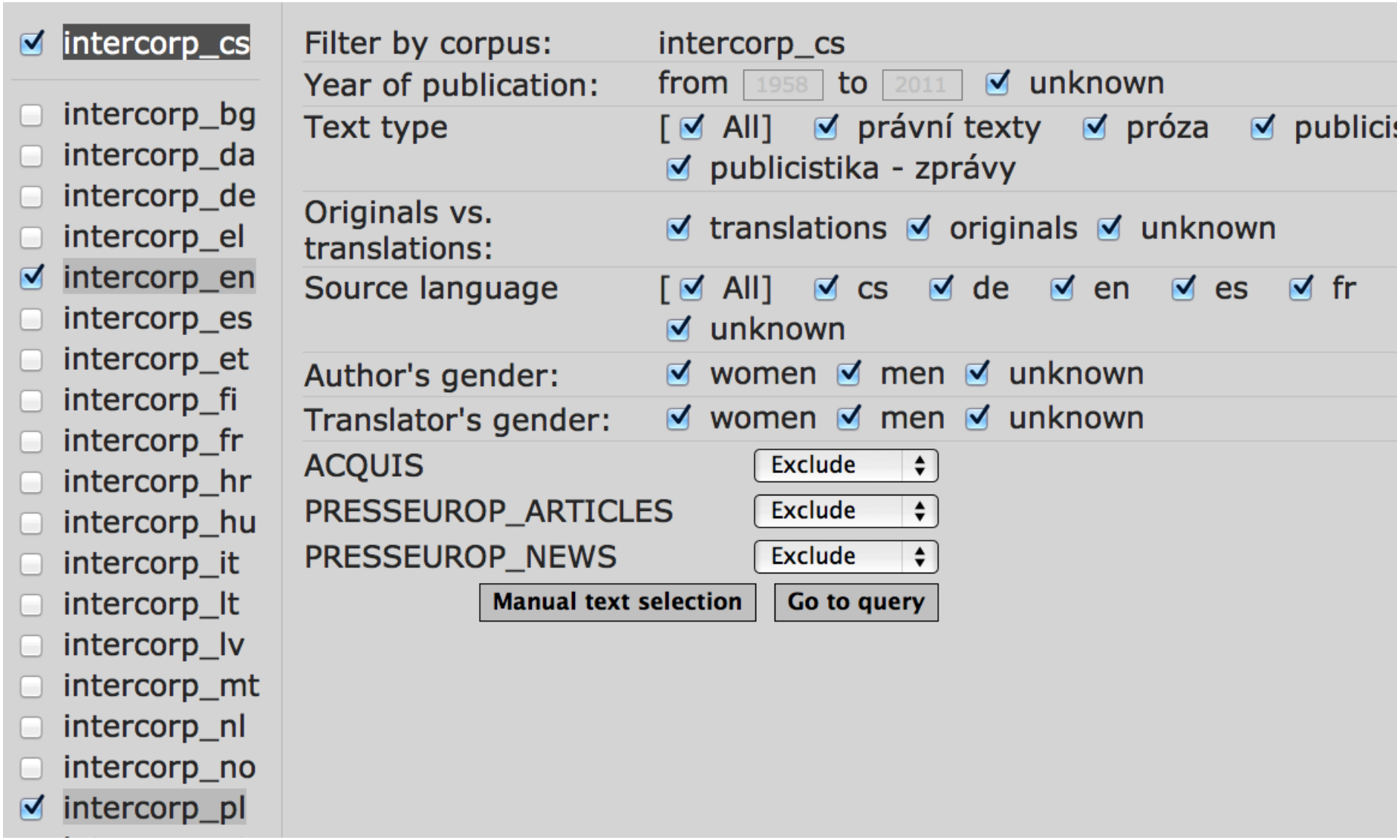
1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Dotazy on-line I

Park

- webové rozhraní pro paralelní korpus
- výběr korpusu vydání 5 nebo 6
- filtry pro výběr textů:
  - jazyky, texty, rok vydání, typ textu
  - originál/překlad, jazyk originálu
  - pohlaví autora a překladatele
- paralelní dotazy, podpora dotazovacího jazyka *Manatee* (jako *Corpus Query Language*)
- pozitivní a negativní filtry konkordancí
- export konkordancí do souboru Excel
- `http://www.korpus.cz/Park/`


Languages
Documents
intercorp_cs
intercorp_be
intercorp_bg
intercorp_da
intercorp_de
intercorp_el
intercorp_en
intercorp_es
intercorp_et
intercorp_fi
intercorp_fr
intercorp_hr
intercorp_hu
intercorp_it
intercorp_lt
intercorp_lv
intercorp_mk
intercorp_mt
intercorp_nl
Filter by corpus:
intercorp_cs
Year of publication:
from 1922 to 2011 unknown
Text type
[All] právní texty próza literatura faktu poe
publicistika - komentáře publicistika - zprávy
Originals vs. translations:
translations originals unknown
Source language
[All] bg cs da de en es fi
no pl pt ro ru sk sl sv ur
Author's gender:
women men unknown
Translator's gender:
women men unknown
ACQUIS
Exclude
PRESSEUROP_ARTICLES
Exclude
PRESSEUROP_NEWS
Exclude
SYNDICATE-2000_2008
Exclude
SYNDICATE-2008_2010
Exclude
Manual text selection
Go to query


intercorp_cs
intercorp_bg
intercorp_da
intercorp_de
intercorp_el
intercorp_en
intercorp_es
intercorp_et
intercorp_fi
intercorp_fr
intercorp_hr
intercorp_hu
intercorp_it
intercorp_lt
intercorp_lv
intercorp_mt
intercorp_nl
intercorp_no
intercorp_pl
Filter by corpus: intercorp_cs
Year of publication: from 1958 to 2011 unknown
Text type [All] právní texty próza publici
publicistika - zprávy
Originals vs. translations: translations originals unknown
Source language [All] cs de en es fr
unknown
Author's gender: women men unknown
Translator's gender: women men unknown
ACQUIS Exclude
PRESSEUROP_ARTICLES Exclude
PRESSEUROP_NEWS Exclude
Manual text selection
Go to query

Languages

- [x] intercorp_cs
- [ ] intercorp_bg
- [ ] intercorp_da
- [ ] intercorp_de
- [ ] intercorp_el
- [x] intercorp_en
- [ ] intercorp_es
- [ ] intercorp_et
- [ ] intercorp_fi
- [ ] intercorp_fr
- [ ] intercorp_hr
- [ ] intercorp_hu
- [ ] intercorp_it
- [ ] intercorp_lt
- [ ] intercorp_lv
- [ ] intercorp_mt
- [ ] intercorp_nl
- [ ] intercorp_no
- [x] intercorp_pl
- [ ] intercorp_pt
- [ ] intercorp_ro
- [ ] intercorp_ru
- [ ] intercorp_sk
- [ ] intercorp_sl
- [ ] intercorp_sr

Documents

Filter by corpus: intercorp_cs

Year of publication: from 1958 to 2011 ☑ unknown

Text type: [☑ All] ☑ právní texty ☑ próza ☑ publicistika - komentáře ☑ publicistika - zprávy

Originals vs. translations: ☑ translations ☑ originals ☑ unknown

Source language: [☑ All] ☑ cs ☑ de ☑ en ☑ es ☑ fr ☑ it ☑ nl ☑ pl ☑ pt ☑ ro ☑ unknown

Author's gender: ☑ women ☑ men ☑ unknown

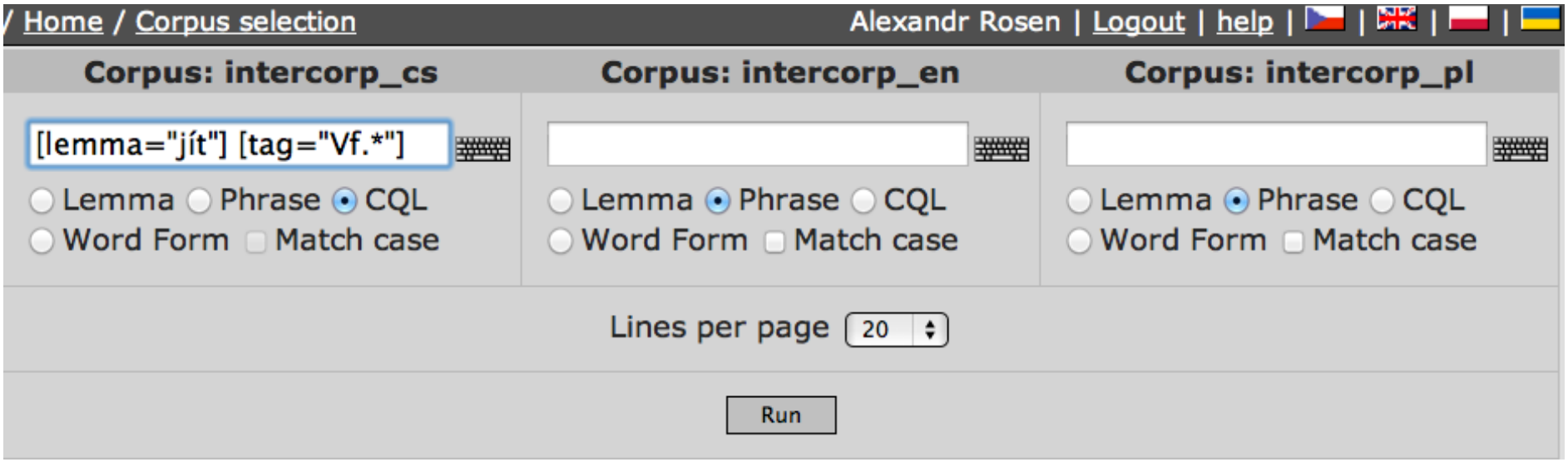
Translator's gender: ☑ women ☑ men ☑ unknown

ACQUIS	Exclude
PRESSEUROP_ARTICLES	Include
PRESSEUROP_NEWS	Include

Filter texts | Manual text selection | Go to query

Check All / Uncheck All

- [x] Grisham-Partner
- [x] Kundera-Nesnesit_lehko
- [x] Orwell-1984
- [x] Viewegh-Vychova
- [x] Kundera-Nesmrtelnost
- [x] kundera-zert
- [x] rowlingova-hpot_kamen


/ Home / Corpus selection
Alexandr Rosen | Logout | help |
Corpus: intercorp_cs
Corpus: intercorp_en
Corpus: intercorp_pl
[lemma="jít"] [tag="Vf.*"]
Lemma Phrase CQL
Word Form Match case
Lemma Phrase CQL
Word Form Match case
Lemma Phrase CQL
Word Form Match case
Lines per page 20
Run

/ Home / Corpus selection / Query | Concordance | Alexandr Rosen | Logout | help | | | | |

1-20 from **71** (page 1 from 4 >>) last | next page >>

intercorp_cs *(539728 tokens)*	intercorp_en *(653269 tokens)*	intercorp_pl *(551869 tokens)*
Kwic		
Možná si šel zaběhat a zanedbal obvyklý postup .	show context Maybe he 'd gone jogging and neglected the routine .	show context Niewykluczone , iż wyruszył na trasę biegu , zapomniawszy dopełnić rutynowych czynności .
O půlnoci si šla lehnout .	show context She went to bed at midnight .	show context Położyła się spać o północy .
Měl jediný pokoj ; dveře se daly zamknout zvenčí a okna nešla otevřít .	show context His was the only room with doors which could be locked from the- outside and windows that would n't open .	show context Patricka umieszczono w niewielkiej salce na samym końcu bocznego skrzydła szpitala . Pojedyncze okno nie dało się otworzyć , drzwi były zamykane od zewnątrz .
Pak pozval své nové kamarády z řad novinářů , aby ho doprovodili pěšky k soudu , až půjde podat žalobu .	show context Then he invited his new pals in the press corps to stroll with him down the sidewalk as he went to file it .	show context A po jej zakończeniu wyruszył wraz z całym tłumem dziennikarzy do kancelarii , aby złożyć pozew .
Paulo také věděl o dost podezřelých individuích , kteří se potulovali u nich v ulici a sledovali ho , když šel nakupovat nebo jel do své pracovny na Pontificia Universidade Católica .	show context Paulo was also tired of the shady little men lurking around his street and following him as he walked to the market or drove to his office at the Pontificia Universidade Catolica .	show context Miał także dosyć ciągłego towarzystwa jakichś podejrzanych typków , którzy nieodmiennie go obserwowali , czy to wychodził na zakupy , czy jechał do swojego gabinetu w Pontificia Universidade Católica .

Export: xls1, xls2 | show context | show context | View: horizontal

Dotazy on-line II

NoSketch Engine

- jednotné prostředí pro hledání v jednojazykových i paralelních korpusech
- brzy by mělo obsahovat všechny funkce *Park*u
- žádný z jazyků nemá privilegované postavení
- v jednotlivých jazycích se dá hledat jako v samostatných korpusech
- více funkcí pro zpracování výsledků dotazu (třídění, frekvenční distribuce, kolokace)
- možnost zobrazení výsledků i v případě, že v některém z jazyků daný text chybí
- texty k prohledávání se vybírají vytvářením subkorpusů.
- `https://korpus.cz/corpora/`

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

Osnova

1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Příprava textů

1. akvizice
2. skenování a rozpoznávání znaků
3. korektury
4. segmentace (zjišťování hranic vět)
5. zarovnání po větách
6. kontrola a opravy segmentace a zarovnání
7. morfosyntaktické značkování

Nástroje pro přípravu textů

1. bibliografická databáze
2. *InterText* – editor paralelních textů
3. *Punkt* – dělič vět
4. *Hunalign* – zarovnávač
5. taggery pro jednotlivé jazyky

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 používají ve vyhledávači

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`

Lingvistické značkování

Strategie pro lingvistické značkování (lemmatizace a morfosyntaktické značkování)

- Používat dostupné nástroje (taggery), včetně:
  - tokenizace (dělení na slova) obsažené v daném nástroji
  - různých sad značek, které vycházejí z různých koncepcí

Současný stav

- Morphosyntaktické značky pro češtinu + 16 cizích jazyků
- Lemmata pro češtinu + 13 cizích jazyků

Nástroje pro lemmatizaci a značkování

Jazyk	Zn.	Lm.	Nástroj	Předl. Det. Adj. Subst.
bg	✓		TT	R Pde-os-n Ansi Ncnsi
cs	✓	✓	Morče	RR-6 PDXP6 AAFP6---3A NNFP6---A
de	✓	✓	TT	APPR ART ADJA NN
en	✓	✓	TT	IN DT JJS NNS
es	✓	✓	TT	PREP ART NC ADJ
et	✓	✓	TT	P--s3 A-p-s3 Nc-s3
fr	✓	✓	TT	PRP DET:ART ADJ NOM
hu	✓		HunPos	ART ADJ ADJ NOUN (CAS (ILL))
it	✓	✓	TT	PRE PRO:demo NOM ADJ
lt	✓	✓	V.D.	prln jvrd bdvr dktv
nl	✓		TT	600 370 103 000
no	✓	✓	OB	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

Osnova

1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Problémy

- více verzí jednoho titulu v jednom jazyku (různé překlady)
  - celý korpus se vlastně skládá ze subkorpusů pro jednotlivé jazyky – není jasné, jak naložit s více verzemi v jednom jazyce
  - není jasné, jak vyhledávat a zobrazovat
- některé jazyky a typy textů nejsou dostatečně zastoupeny
- některé jazyky ještě nemají lemmata a značky
- odlišná pravidla tokenizace a sady značek pro různé jazyky

Problémy se značkováním

Důsledky oportunistické strategie:

- odlišná pradidla tokenizace, např. pro spřežky a spojovník
  abychom, udělals, tys, očs
  ×
  że+by+śmy, zrobił+eś, ty+ś, gdzieś/gdzie+ś
  ca+n't, I+'m

- odlišné zacházení s víceslovnými výrazy
  Estados~Unidos (NP), *a~lo~largo~de* (PREP), *tendrán~que* (VMfin), *por~el~momento* (ADV), *al~mismo~tiempo* (ADV)

- odlišné sady značek

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

Osnova

1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Využití korpusu

- vylepšování vyhledávacího rozhraní
- integrace s jinými paralelními korpusy?

Obsah

- lepší rovnováha mezi jazyky a typy textů
- více jazyků: albánština, čínština, romština, vietnamština, lužická srbština ?

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.
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.
Oikein paljon kiitoksia mielenkiinnostanne.
Je vous remercie de votre attention.
Nagyon szépen köszönöm a figyelmüket.
Velice vám děkuji za pozornost.

Osnova

1. O korpusu InterCorpu
   - Základní údaje
   - Obsah korpusu
2. Některé podobné korpusy
3. Jak korpus využívat
   - Dotazy on-line
   - Poskytování úplných textů
4. Příprava textů
   - Bibliografická databáze
   - Zarovnání
   - Lingvistické značkování
5. Problémy
   - Problémy se značkováním
6. Perspektivy
7. Dodatky
   - Zastoupení slovanských jazyků
   - Údaje o využívání korpusu InterCorp

Slovanské jazyky v InterCorpu – počet slov

	Jazyk	Jádro	Syndicate	PressEU	Acquis	Europarl	Celkem
be	běloruština	1 308	0	0	0	0	1 308
bg	bulharština	3 979	0	0	13 816	9 083	26 879
cs	čeština	61 962	2 741	1 639	20 285	12 920	99 547
hr	chorvatština	**12 625**	0	0	0	0	12 625
mk	makedonština	2 664	0	0	0	0	2 664
pl	polština	**12 710**	0	1 660	20 464	12 805	47 640
ru	ruština	4 937	2 651	0	0	0	7 588
sk	slovenština	**8 152**	0	0	19 222	12 734	40 108
sl	slovinština	1 855	0	0	19 646	12 241	33 741
sr	srbština	**6 972**	0	0	0	0	6 972
uk	ukrajinština	1 493	0	0	0	0	1 493
de	němčina	17 256	3 050	1 715	21 724	13 089	56 835
en	angličtina	10 019	3 083	1 863	24 208	15 580	54 753
es	španělština	14 552	3 479	1 948	27 001	15 885	62 865
fr	francouzština	3 816	3 535	2 054	27 352	17 178	53 936

Slovanské jazyky v InterCorpu – počet textů

Jazyk	Celkem	Originálů	Originálů v jiném slovanském jazyce
be	28	0	12
bg	54	19	7
hr	188	22	63
mk	29	0	6
pl	180	41	40
ru	104	22	11
sk	155	55	62
sl	31	2	10
sr	86	0	23
uk	26	0	6
Celkem	**881**	**161**	

Tituly ve všech slovanských jazycích

- Michail Bulgakov *Mistr a Markétka*
- Nikolaj Ostrovskij *Jak se kalila ocel*
- Lewis Carroll *Alenka v říši divů*
- Alan Milne *Medvídek Pú*
- J. K. Rowling *Harry Potter a kámen mudrců*
- Antoine de Saint Exupery *Malý princ*
- J. R. R. Tolkien *Hobit aneb cesta tam a zase zpátky*

Údaje o využívání korpusu InterCorp

- za leden – říjen 2012
- podle jazyků
- počítá se každé kliknutí na příslušný jazyk

	01	02	03	04	05	06	07	08	09	10
be	0	0	0	0	0	1	3	10	0	7
bg	103	16	16	8	22	102	77	111	301	37
da	0	12	8	15	151	1	23	3	1	3
de	1183	675	1249	1314	799	1155	972	2104	834	1592
el	0	0	0	0	0	1	2	0	0	1
en	689	800	1011	1611	1264	936	840	1197	886	2090
es	14	98	222	67	246	68	58	8	12	88
et	0	0	0	0	0	1	0	0	0	1
fi	14	54	28	235	437	3	6	2	33	5
fr	90	715	1142	1661	1737	488	320	171	300	957
hr	4	41	120	0	33	33	73	76	46	17
hu	0	1	12	6	22	7	0	2	0	3
it	179	48	538	421	204	733	135	524	222	297
lt	0	2	57	6	3	8	2	3	3	13
lv	7	3	45	1	16	10	10	15	6	2
mk	0	0	0	0	0	0	0	0	7	2
mt	0	0	0	0	0	2	0	0	0	0
nl	6	67	11	9	2	76	0	8	0	113
no	26	110	5	11	21	2	0	0	0	29
pl	102	37	220	111	256	55	76	364	24	684
pt	14	97	290	202	38	2	30	0	7	1
ro	6	0	1	1	1	44	0	0	1	0
ru	202	61	117	213	216	57	61	379	80	52
sk	9	7	8	4	33	7	14	10	37	41
sl	0	2	74	10	24	23	3	7	13	5
sr	0	11	29	4	5	9	3	67	36	88
sv	6	2	13	15	11	8	2	66	11	59