Paralelní korpusy – úvod

Seminář ÚČNK, 3. března 2016

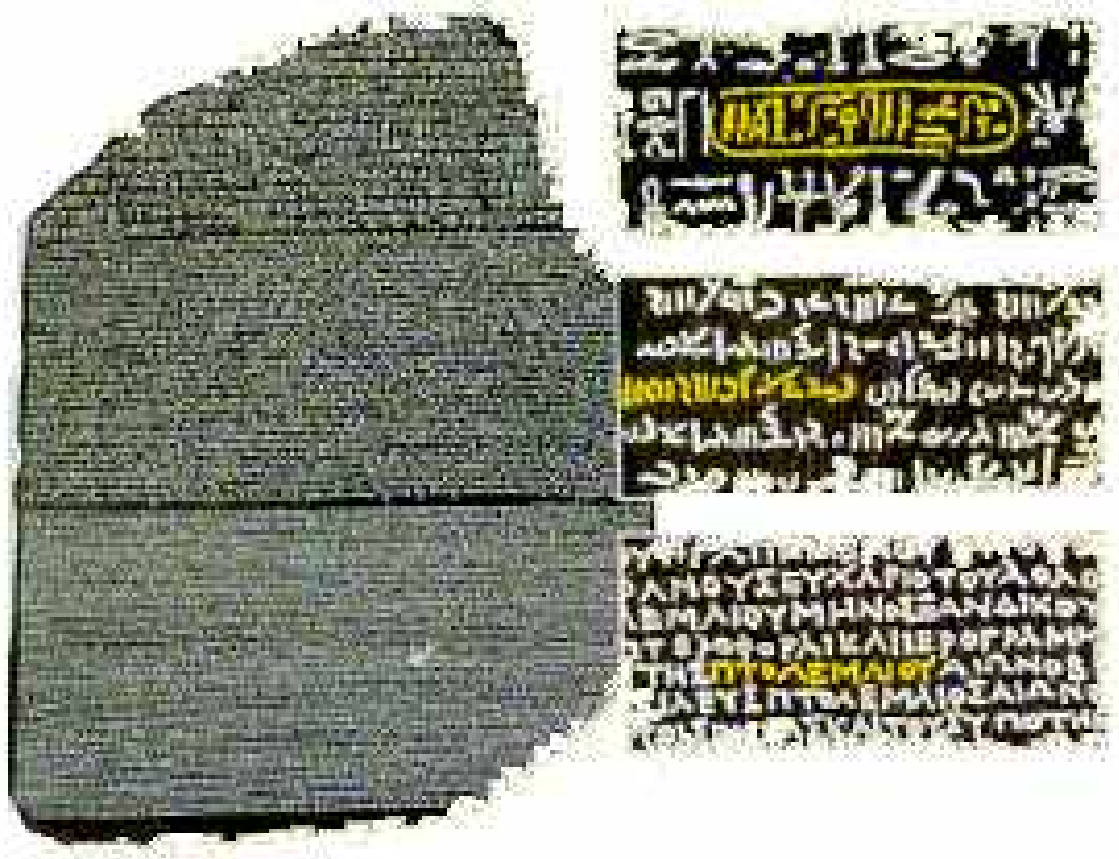
Alexandr Rosen

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/pc2016.pdf

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

Co je to paralelní korpus?

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



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

Podmínky pro rozumnou práci s paralelními korpusy:

- zarovnání po větách
- paralelní korpusový manažer (*concordancer*)

Nevýhody paralelních korpusů:

- texty nejsou autentické, většinou jen překlady
- texty nejsou reprezentativní,
  paralelně lze získat jen některé typy textů
- předpokladem rozumného využití je spolehlivé zarovnání po větách, ale:
  - zarovnávat ručně je dřina
  - automatické metody zarovnávání nefungují na 100 %
- není snadné získat nástroje, které mají požadované funkce a přitom nevyžadují speciální znalosti

- 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:

- statistický strojový překlad
  (*Statistical Machine Translation*)
- strojový překlad podle příkladů
  (*Example-based 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.

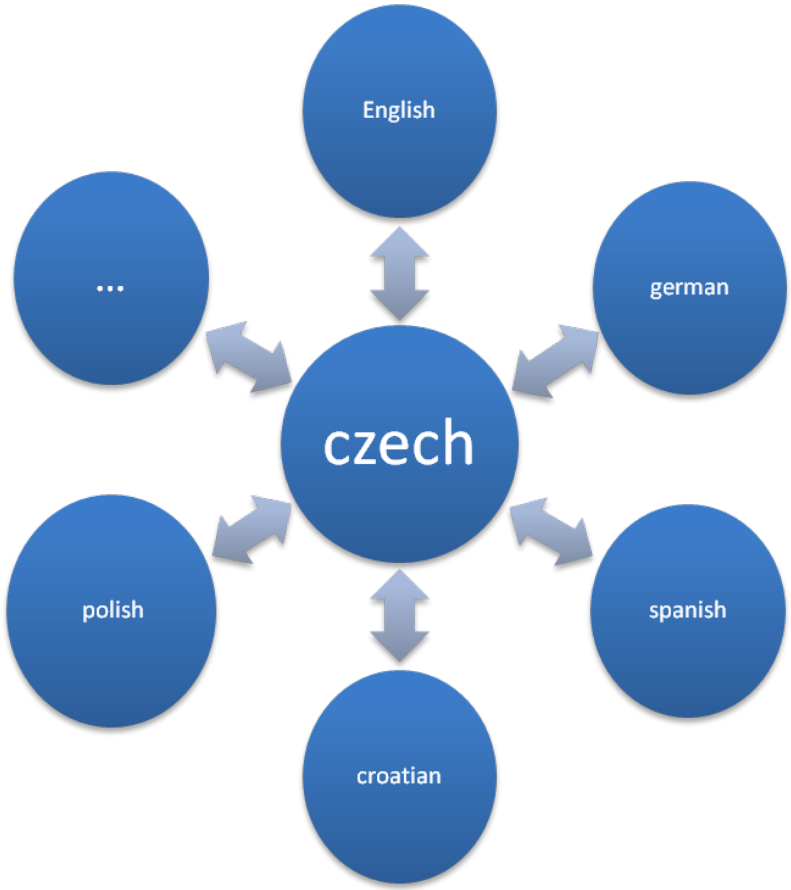
- Základní údaje
- Obsah korpusu

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, ...)

- Základní údaje
- Obsah korpusu

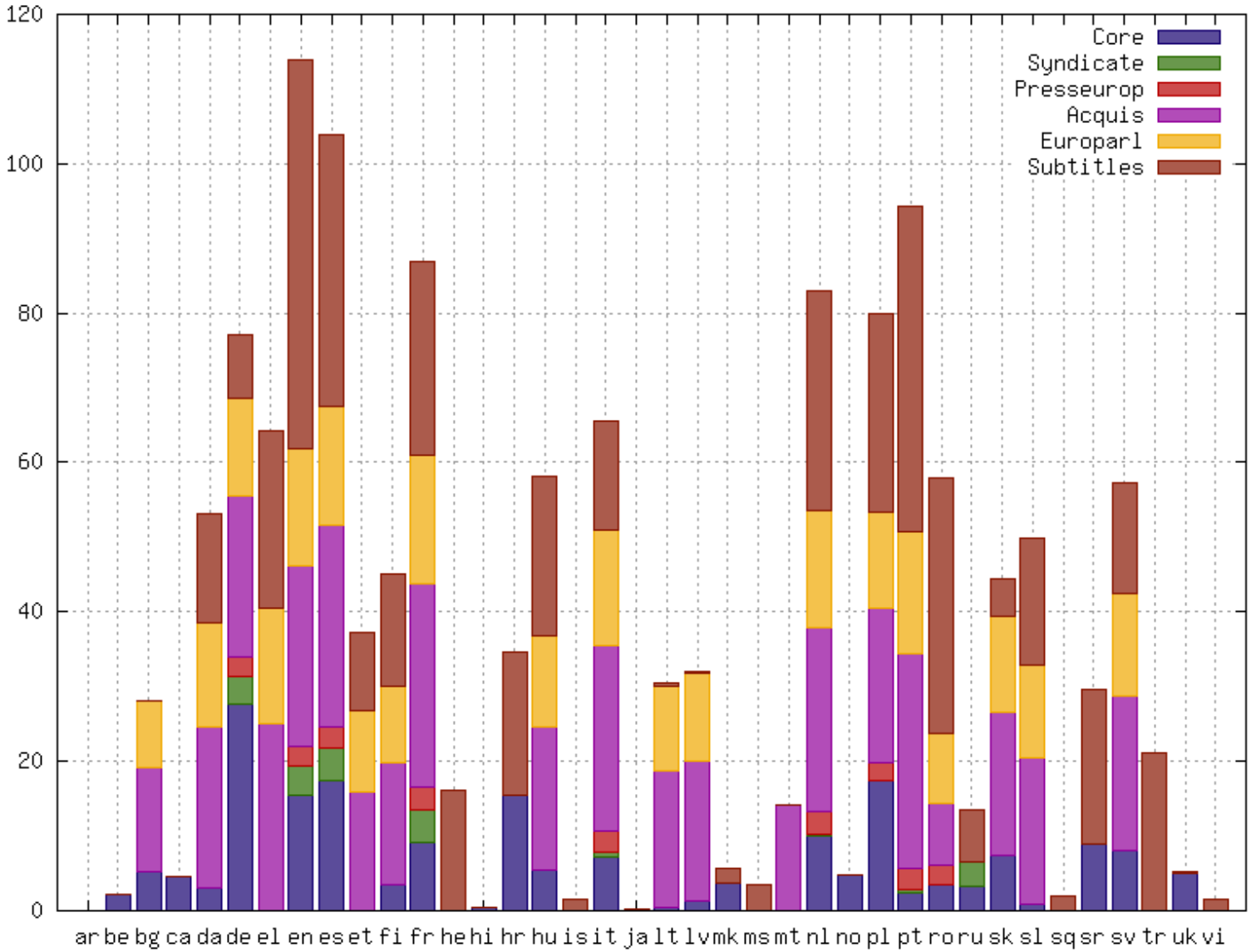
Vývoj

verze	rok	mil. cizích slov	cizích jazyků	s tagy
8	2015	1 423,0	38	20
7	2014	1 390,0	38	20
6	2013	867,3	31	17
5	2012	542,6	27	17
4	2011	92,3	22	13
3	2011	72,3	22	13
2	2009	49,3	21	10
1	2009	34,5	20	10
0	2008	25,0	19	0

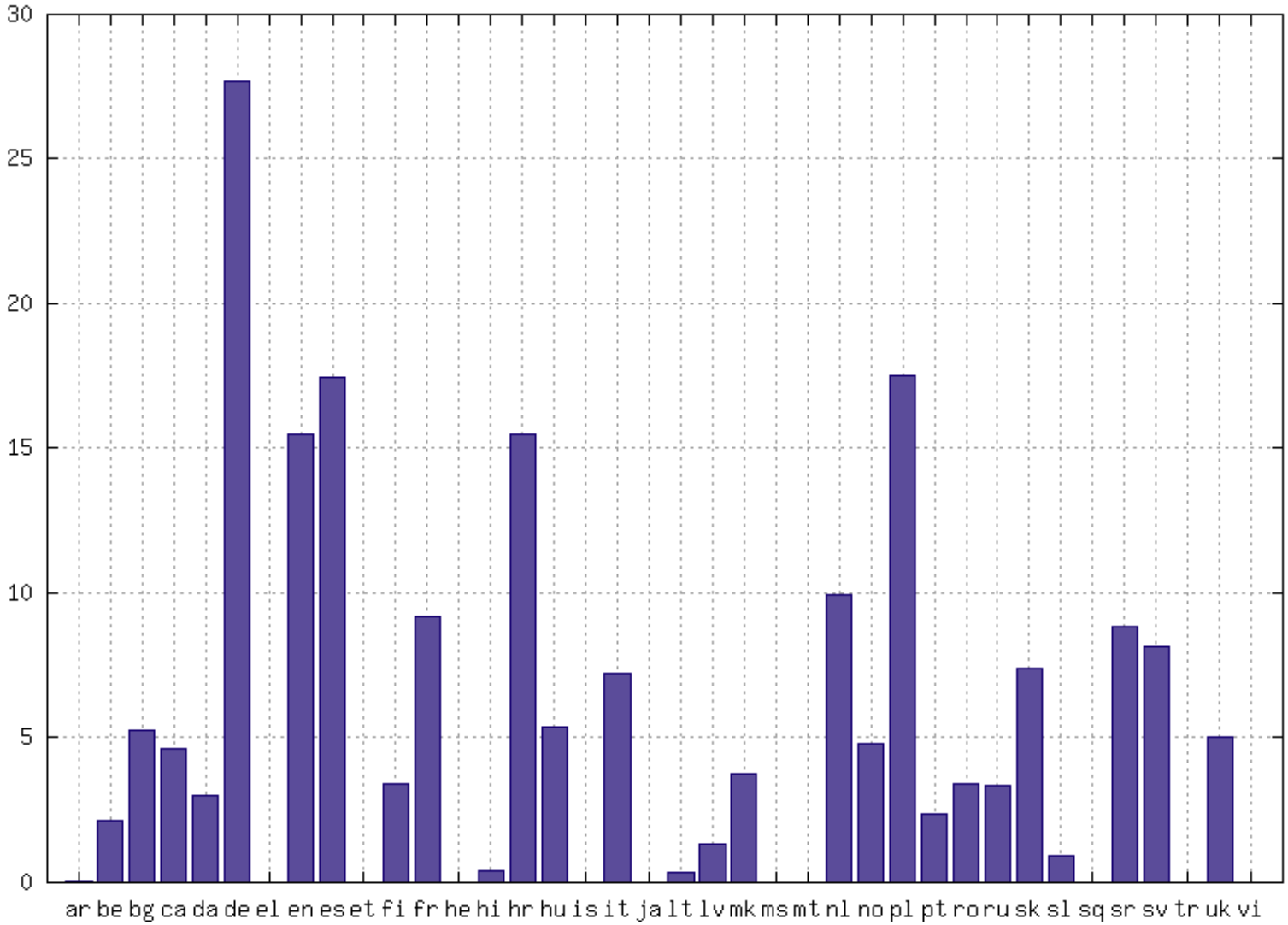
Obsah korpusu – 8. vydání

- **Počet jazyků**: 38 + česky
  - jen málo textů je k mání ve více než 5 jazycích
  - jazyky se velmi liší objemem textů
- **Celková velikost** – 1794/232 mil. slov (cizí/české)
- **Jádro** – 194/84 mil. slov: beletrie s manuálně zkorigovaným zarovnáním
- **Kolekce** – texty s automatickým zarovnáním:
  - **Žurnalistika** – 45/6 mil. slov:
    Project Syndicate http://www.project-syndicate.org/
    Voxeurop http://www.voxeurop.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/
  - **Filmové titulky** – 488/51 mil. slov:
    Open Subtitles http://www.opensubtitles.org

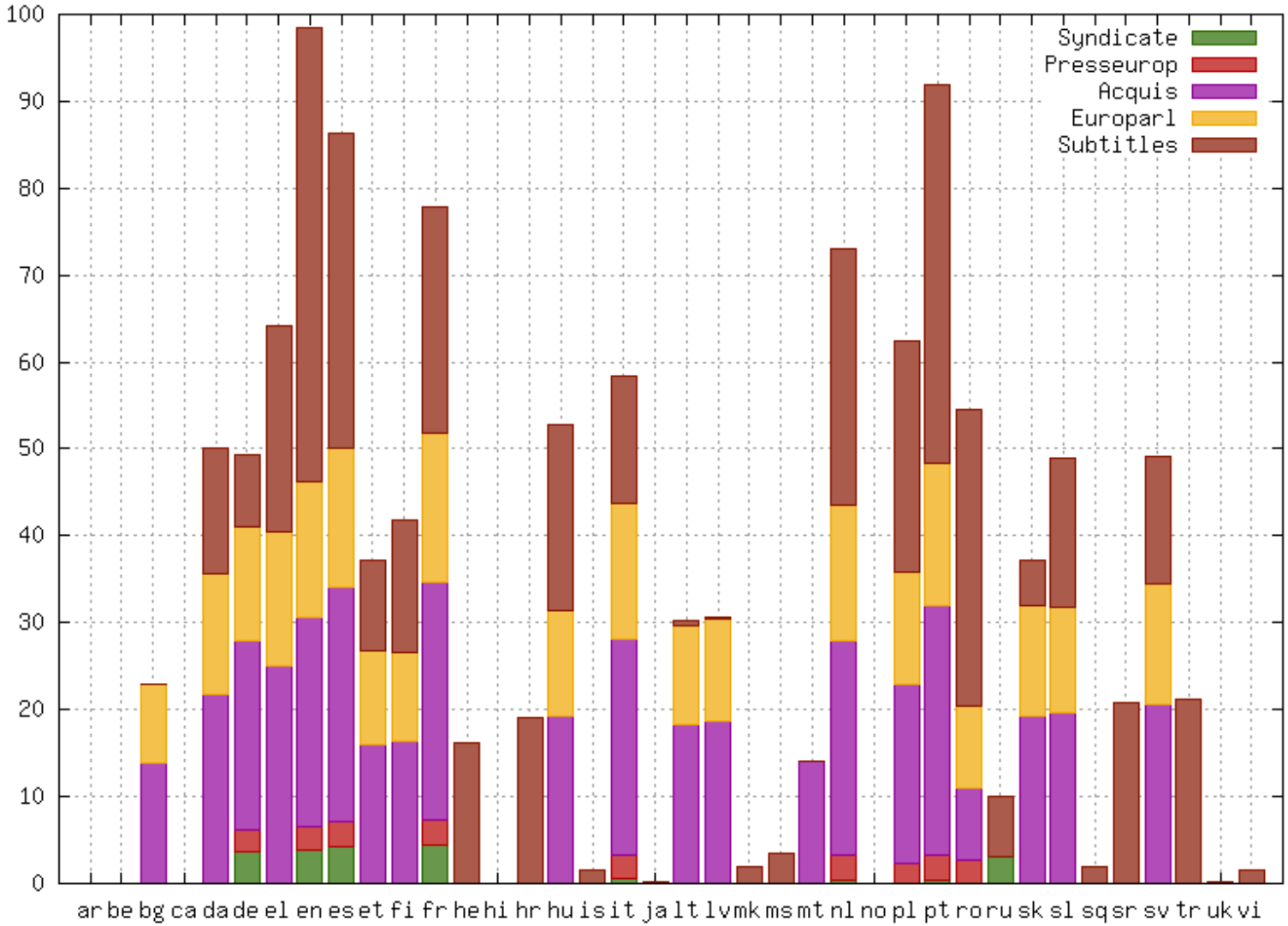
Obsah korpusu podle jazyků a typu textů



Jádro (beletrie)



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



27 titulů s nejvíce překlady (1/2)

26	J. K. Rowlingová *Harry Potter a kámen mudrců*
26	A. de Saint-Exupéry *Malý princ*
23	Lewis Carroll *Alenka v říši divů*
21	Milan Kundera *Nesnesitelná lehkost bytí*
21	J. K. Rowlingová *Harry Potter a tajemná komnata*
21	J. R. R. Tolkien *Společenství prstenu*
20	Douglas Adams *Stopařův průvodce po galaxii*
20	Milan Kundera *Žert*
20	J. R. R. Tolkien *Návrat krále*
19	Dan Brown *Šifra Mistra Leonarda*
19	Michail Bulgakov *Mistr a Markétka*
19	J. R. R. Tolkien *Dvě věže*
19	J. K. Rowlingová *Harry Potter a vězeň z Azkabanu*
18	Umberto Eco *Jméno růže*

27 titulů s nejvíce překlady (2/2)

18	Jaroslav Hašek *Osudy dobrého vojáka Švejka*
18	A. A. Milne *Medvídek Pú*
18	J. R. R. Tolkien *Hobit*
17	Paolo Coelho *Alchymista*
17	Franz Kafka *Proces*
17	George Orwell *1984*
17	J. K. Rowling *Harry Potter a ohnivý pohár*
16	Anna Franková *Deník*
16	Rudyard Kipling *Kniha džunglí*
16	Milan Kundera *Nesmrtelnost*
16	Bohumil Hrabal *Obsluhoval jsem anglického krále*
15	Milan Kundera *Směšné lásky*
15	J. K. Rowling *Harry Potter a Fénixův řád*

Některé podobné korpusy

Název	Texty	Jaz.	Velikost	Anot.	Zar.	Man.	Hled.	Staž.	Meta
Linguee	práv.	25	?	–	V,S	–	+	–	+
Glosbe	růz.	100+	1MldV	–	V,S	–	+	–	+
Tatoeba	růz.	?	?	–	V	+	+	+	–
SKE	růz.	38	cs:217MS	–	V	–	+	+	+
DGT-TM	práv.	22	cs:3.7MS	–	V	+	–	+	–
Pelcra	růz.	31	pl:58MS	–	V,S	část	+	+	+
RNC	růz.	6	9MS	M	V	část	+	?	+
SNK	bel.	7	sk:388MS	M	V	–	+	část	+
CzEng	růz.	en,cs	en:233MS	M,S	V	–	+	+	–
PCEDT	nov.	en,cs	1.2MS	M,S	V,S	+	+	+	+
Kačenka	bel.	en,cs	3.3MS	–	V	+	–	+	+
Opus	růz.	100+	4.7MldS	M,S	V,S	–	+	+	–
Parasol	bel.	31	27MS	M	V	část	+	?	+
ASPAC	bel.	25	68 textů	–	O	+	–	?	+
InterCorp	růz.	32	1.6MldS	M	V	část	+	+	+

Links 1/2

- Linguee: Online search through bilingual texts – http://www.linguee.com
- Glosbe: Translation Memory Online – http://glosbe.com/tmem/
- Tatoeba: Collection of sentences and translations – https://tatoeba.org/eng/
- SKE: Sketch Engine – http://www.sketchengine.co.uk
- DGT-TM: Translation Memory of the EC'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

Links 2/2

- 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
- Opus: An open source parallel corpus – http://opus.lingfil.uu.se
- Parasol: A Parallel Corpus of Slavic and other languages – http://www.slavist.de
- ASPAC: The Amsterdam Slavic Parallel Corpus – http://home.medewerker.uva.nl/a.a.barentsen/

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 (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

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

- Dotazy on-line
- Poskytování úplných textů
- Statistika přístupů

Dotazy on-line

KonText

- jednotné prostředí pro hledání v jednojazykových i paralelních korpusech
- 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
- https://kontext.korpus.cz

treq – Lexikální ekvivalenty podle zarovnání po slovech

- https://trnka.ff.cuni.cz/~vavrin/ic_slovnik/index.php

- Dotazy on-line
- Poskytování úplných textů
- Statistika přístupů

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

- Dotazy on-line
- Poskytování úplných textů
- Statistika přístupů

Statistika přístupů

- Leden–červen 2015
- Počet dotazů na daný jazyk a typ textu v libovolné kombinaci s jinými jazyky
- Celkem počet dotazů na kombinaci jazyků: 62 tisíc, z toho 50 tisíc na dva jazyky, 10 tisíc na jediný jazyk

	-	Acq	Acq/Parl	Core	Parl	Vox	Tit	Synd	Kol	TOTAL	%
ar	122	0	0	6	1	0	1	0	0	130	0.11%
be	10	0	0	1	0	0	0	0	0	11	0.01%
bg	418	0	0	9	0	0	0	0	1	428	0.37%
ca	10	0	0	0	0	0	0	0	0	10	0.01%
cs	46,220	215	91	2,871	248	90	467	341	76	50,619	44.33%
da	10	0	0	4	0	0	0	0	0	14	0.01%
de	10,256	8	0	106	14	8	28	80	42	10,542	9.23%
el	373	0	0	0	0	0	0	0	0	373	0.33%
en	20,368	117	1	1,402	189	26	177	88	12	22,380	19.60%
es	8,920	37	0	440	10	7	2	51	30	9,497	8.32%
et	1	0	0	0	0	0	0	0	0	1	0.00%
fi	1,871	7	0	178	3	0	103	0	0	2,162	1.89%
fr	4,504	39	0	392	36	93	237	286	10	5,597	4.90%
he	5	0	0	0	0	0	0	0	0	5	0.00%
hi	8	0	0	0	0	0	0	0	0	8	0.01%
hr	262	0	0	22	0	0	2	0	0	286	0.25%
hu	144	0	0	2	0	0	1	0	0	147	0.13%
is	14	0	0	0	0	0	0	0	0	14	0.01%
it	1,740	44	0	260	11	7	37	27	0	2,126	1.86%
ja	42	0	0	0	0	0	0	0	0	42	0.04%
lt	138	0	0	0	0	0	0	0	0	138	0.12%
lv	411	0	0	1	0	0	0	0	0	412	0.36%
mk	50	0	0	5	0	0	0	0	0	55	0.05%
nl	2,503	0	0	412	0	1	30	3	0	2,949	2.58%
no	21	0	0	1	0	0	0	0	0	22	0.02%
pl	1,644	6	0	192	7	7	42	13	11	1,922	1.68%
pt	240	4	0	12	0	1	1	0	0	258	0.23%
ro	7	0	0	0	0	0	0	0	0	7	0.01%
ru	2,076	0	0	52	0	0	8	10	1	2,147	1.88%
sk	634	0	0	21	2	0	2	0	11	670	0.59%
sl	66	0	0	1	0	0	0	0	0	67	0.06%
sq	6	0	0	0	0	0	0	0	0	6	0.01%
sr	73	0	0	14	0	0	0	0	0	87	0.08%
sv	753	0	0	8	8	1	8	2	0	780	0.68%
sy	19	0	0	0	0	0	0	0	0	19	0.02%
tr	118	0	0	0	0	0	0	0	0	118	0.10%
uk	94	0	0	4	0	0	2	13	0	113	0.10%
vi	10	0	0	0	0	0	0	1	0	11	0.01%
TOTAL	104,169	477	92	6,416	529	241	1,148	915	194	114,181	100.00%
%	91.23%	0.42%	0.08%	5.62%	0.46%	0.21%	1.01%	0.80%	0.17%	100.00%	

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
2. *Intertext* – editor paralelních textů
3. *Punkt* – větný segmentátor
4. *Hunalign* – zarovnávač
5. taggery pro některé jazyky

- Bibliografická databáze
- Zarovnání
- Lingvistické značkování

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

- Bibliografická databáze
- Zarovnání
- Lingvistické značkování

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?

- Bibliografická databáze
- Zarovnání
- Lingvistické značkování

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 + 19 cizích jazyků
- Lemmata pro češtinu + 16 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

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

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

Perspektivy

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

Díky za pozornost!


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