

# Paralelní korpusy – úvod

Seminář ÚČNK, 12. března 2015

Alexandr Rosen

Ústav teoretické a počítačové lingvistiky  
Filozofické fakulty Univerzity Karlovy v Praze  
[alexandr.rosen@ff.cuni.cz](mailto:alexandr.rosen@ff.cuni.cz)  
<http://utkl.ff.cuni.cz/~rosen>

- 1 Úvod
  - Korpusy a paralelní korpusy
  - K čemu je paralelní korpus?
  - Ukázky paralelních konkordancí
- 2 O InterCorpu
  - Základní údaje
  - Obsah korpusu
- 3 Některé podobné korpusy
- 4 Jak korpus používat
  - Dotazy on-line
  - Poskytování úplných textů
  - Statistika přístupů
- 5 Příprava textů
  - Bibliografická databáze
  - Zarovnání
  - Lingvistické značkování
- 6 Problémy a perspektivy

1 Úvod

2 O InterCorpu

3 Některé podobné korpusy

4 Jak korpus používat

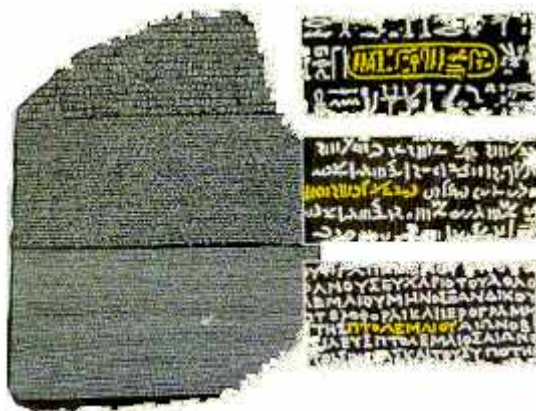
5 Příprava textů

6 Problémy a perspektivy

- **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 vymežit, 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 více-stopě 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 Odone, both **determined**, strong-willed people, had decided they would learn all they could about the disease.

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

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

Jen mu neudělat radost, jen se ne-rozbrečet!

# 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 řídicím 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.

1 Úvod

**2 O InterCorpu**

3 Některé podobné korpusy

4 Jak korpus používat

5 Příprava textů

6 Problémy a perspektivy

- 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

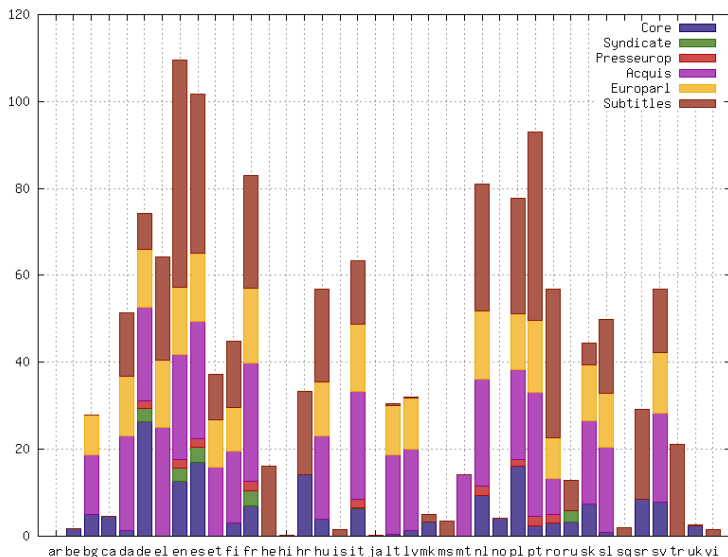
- Získávání a příprava textů:
  - Univerzita Karlova v Praze
  - Masarykova Univerzita v Brně
  - Univerzita Palackého v Olomouci
  - Česká akademie věd
  - Varšavská univerzita
- Pomoc ze zahraničí:
  - texty (ASPAC, Parasol, OPUS, ...)
  - nástroje pro lingvistickou anotaci (TreeTagger, ...)
  - obecnější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

- Základní údaje
- **Obsah korpusu**

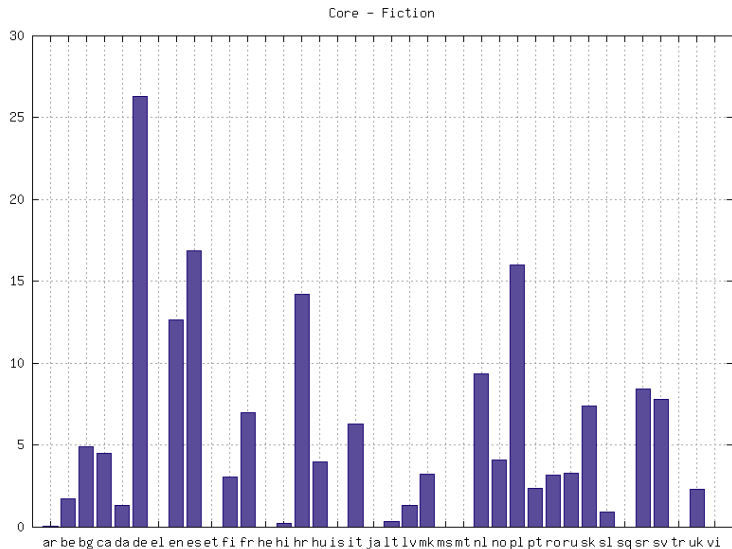
# Obsah korpusu – 7. vydání

- **Počet jazyků:** 38 + česky
  - jen málo textů je k mání ve více než 5 jazycích
  - jazyky se velmi liší objemem textů
- **Celková velikost** – 1390/165 mil. slov (cizí/české)
- **Jádro** – 173/77 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/>
  - **Právníkové 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

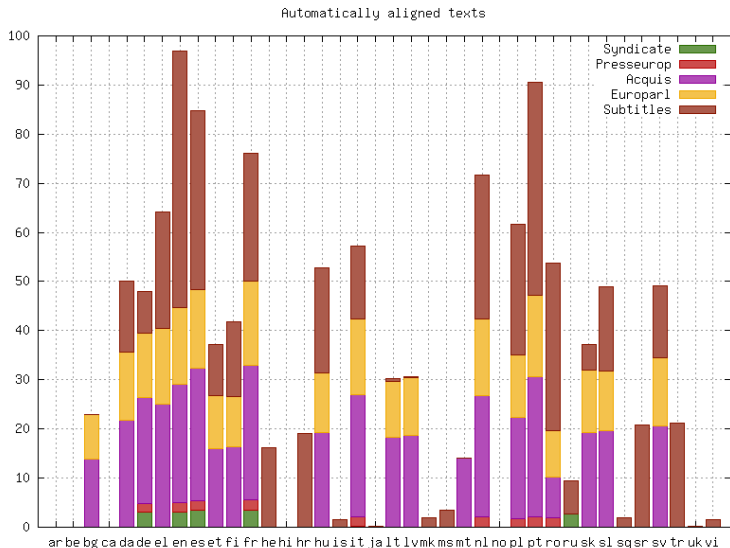


# Jádro (beletrie)





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



## Tituly s nejvyšším počtem verzí I

- 26 J. K. Rowlingová *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. Rowlingová *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. Rowlingová *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*

- 1 Úvod
- 2 O InterCorpu
- 3 Některé podobné korpusy**
- 4 Jak korpus používat
- 5 Příprava textů
- 6 Problémy a perspektivy

## 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 (EMA)  
(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/kacenska/kachna.html>
- PCEDT: Prague Czech-English Dependency Treebank  
[http://ufal.mff.cuni.cz/pcedt/doc/PCEDT\\_main.htm](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/](http://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

- 1 Úvod
- 2 O InterCorpu
- 3 Některé podobné korpusy
- 4 Jak korpus používat**
- 5 Příprava textů
- 6 Problémy a perspektivy

- **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`

# treeq – Lexikální ekvivalenty podle zarovnání po slovech

- [https://trnka.ff.cuni.cz/~vavrin/ic\\_slovník/index.php](https://trnka.ff.cuni.cz/~vavrin/ic_slovník/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

- Za rok 2012
- Podle jazyků
- Počítá se každé kliknutí na 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	26	1	4	9	2	67	36	87
sv	6	2	13	15	11	8	2	66	11	59
sy	0	0	3	3	1	0	1	0	0	1

- 1 Úvod
- 2 O InterCorpu
- 3 Některé podobné korpusy
- 4 Jak korpus používat
- 5 Příprava textů**
- 6 Problémy a perspektivy

## 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>

151	☆	<p>▶ Když to povázím, strýc císaře pána, a voni ho zastřelejí.</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Wenn ich mir das so überleg, ein Onkel Seiner Majestät des Kaisers, und sie erschießen ihn!</p>	<p>✓</p>
152	☆	<p>▶ Vždyť je to ostuda, jsou toho plný noviny.</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Das is ja ein Skandal, die ganzen Zeitungen sind voll damit.</p>	<p>✓</p>
153	☆	<p>▶ U nás před léty v Budějovicích probodli na trhu v nějaké takové malé hádce jednoho obchodníka s dobytkem, nějakého Břetislava Ludvíka. ☹️</p> <p>▶ Ten měl syna Bohuslava, a kam přišel prodávat prasata, nikdo od něho nic nekoupil a každý říkal: ☹️</p> <p>▶ To je syn toho probodnutýho, to bude asi také pěcknej lump.'</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Bei uns in Budweis hat man vor Jahren auf dem Markt bei irgendeinem kleinen Streit einen Viehhändler erstochen, einen gewissen Břetislav 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.“</p>	<p>✓</p>
154	☆	<p>▶ 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."</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ 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."</p>	<p>✓</p>
155	☆	<p>▶ "Vy ale máte divná přirovnání," řekl Bretschneider významně, "mluvíte napřed o Ferdinandovi a potom o obchodníku s dobytkem."</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ „Sie ziehen aber merkwürdige Vergleiche“, sagte Bretschneider bedeutungsvoll, „zuerst sprechen Sie von Ferdinand und dann von einem Viehhändler.“</p>	<p>✓</p>
156	☆	<p>▶ "Ale nemám," hájil se Švejka, "bůh mě chraň, abych já chtěl někoho k někomu přirovnávat."</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ „I wo“, verteidigte sich Schwejk. „Gott bewahre, daß ich jemand mit jemandem vergleichen möchte.“</p>	<p>✓</p>
157	☆	<p>▶ Pan hostinský mne zná.</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Der Herr Wirt kennt mich.</p>	<p>✓</p>
158	☆	<p>▶ Vid' že jsem nikdy nikoho k někomu nepřirovnával?</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Nicht wahr, ich hab nie jemanden mit jemandem verglichen?</p>	<p>✓</p>
159	☆	<p>▶ Já bych jenom nechtěl být v kůži té vdovy po arcivévodovi.</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Ich möchte nur nicht in der Haut der Frau Erzherzogin stecken.</p>	<p>✓</p>
160	☆	<p>▶ Co teď bude dělat?</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▲ ▲ ▲ ▼ ▼ ▼</p>	<p>▶ Was wird die jetzt machen?</p>	<p>✓</p>

- 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-1 P---pl Afp-plf Ncmpln
sk	✓	✓	Morče	Eu6 PFfs6 AAfs6x SSfs6
sl	✓	✓	totale	S1 Pd-nsg Agpfsg Ncns1

- 1 Úvod
- 2 O InterCorpu
- 3 Některé podobné korpusy
- 4 Jak korpus používat
- 5 Příprava textů
- 6 Problémy a perspektivy**

## 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

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

- české **těch** se značuje jako **ukazovací zájmeno**, přičemž se nerozlišuje, zda je užito v pozici substantivní nebo adjektivní
- polské **tym** se značuje 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**.

 Tiedemann, J. (2012).

Parallel data, tools and interfaces in OPUS.

In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

 von Waldenfels, R. (2006).

Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment.

In B. Brehmer, V. Zdanova, and R. Zimny, editors, *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, volume 9, pages 123–138. Verlag Otto Sagner, München.