Paralelní korpusy

0/2 Z, zimní semestr 2006/2007

Alexandr Rosen

Ústav teoretické a komputační lingvistiky
Filozofická fakulta Univerzity Karlovy v Praze

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Abstrakt

- úvodní, prakticky orientovaný kurs
- příprava a využití paralelních korpusů
- DIY: vlastní paralelní korpus!

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Osnova

1. Úvod: korpusy a korpusová lingvistika, paralelní korpusy a jejich využití
2. Ukázky: existující projekty a zdroje dat
3. Výběr a získávání textů: vyváženost korpusu, technické a právní problémy
4. Technické aspekty: formát dat, programové nástroje, hardware
5. Příprava textů: opravy a úpravy, konverze
6. Zarovnávání (alignment): automatické nástroje, kontrola a opravy
7. Hledání v paralelním korpusu: nástroje a práce s nimi
8. Další způsoby využití paralelních korpusů: komputační lexikografie, hledání v cizojazyčných textech, strojový nebo počítačem podporovaný překlad, ...
9. Konzultace k individuálním projektům, jejich prezentace

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Zápočet

- „projekt“
  - individuální nebo skupinový
  - skupina = 2 osoby, výjimky v odůvodněných případech
- náměty:
  - vytvoření paralelního korpusu
  - využití paralelního korpusu

Komunikace

- `http://utkl.ff.cuni.cz/~rosen/VYUKA/MT/pc.html`
- `alexandr.rosen@ff.cuni.cz`
- konzultace v úterý 10:00–12:30, Celetná 13, č. 21, nebo po dohodě
- telefon 221619752, 721451239

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

morfologický přís. jaz.

morfologická rovina

Na základe vzťahov medzi jednotlivými rovinami možno usúdit, že morfologická rovina je v jazykovej stavbe medzi lexikálnou rovinou a syntaktickou rovinou. Gramatické tvary sú totiž pomenovaniami vo vzťahov a sú prostriedkom na vyjadrenie syntaktických funkcií. Toto umiestnenie morfologickej roviny dá sa doložit radom faktov synchronickej i diachronickej povahy. .... z uvedených troch rovín jazykového systému je lexikálna rovina a syntaktická rovina prvotná (prius), kým morfologická rovina vzhľadom na obidve je druhotná (posterius).

%/.

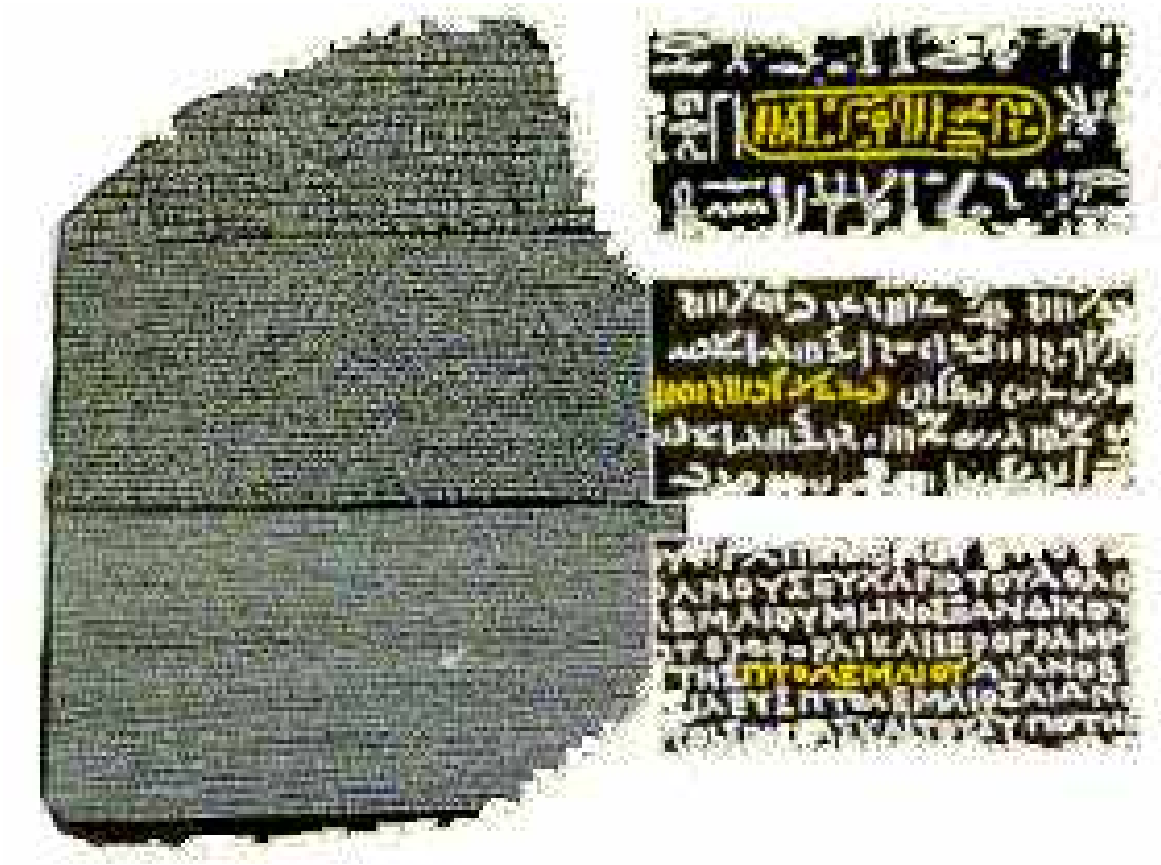
Co je to korpus?

- rozsáhlý soubor elektronicky uložených jazykových dat určený k vědeckému výzkumu (*Encyklopedie Diderot*)
- soubor jazykových (analyzovaných a vykládaných) materiálů (vět, textů ap.) (*SSJČ*)
- soubor počítačově uložených textů, který slouží k výzkumu jazyka, k práci s korpusy se používají speciální programy, které umožňují vyhledávání slov a slovních spojení v kontextu, zjištění frekvence výskytu v korpusu i zjištění původního zdroje textu (*Wikipedia*)
- vnitřně strukturovaný, unifikovaný a obvykle i oindexovaný a ucelený rozsáhlý souhrn elektronicky uložených a zpracovávaných jazykových dat většinou v textové podobě, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž je také považován za reprezentativní (*F. Čermák: Jazykový korpus – prostředek a zdroj poznání, SaS 1995*)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Co je to paralelní korpus?

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



Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Typy paralelních korpusů:

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

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

Podmínky pro rozumnou práci s paralelními korpusy:

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

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

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 automatické metody zarovnávání nefungují na 100 %
- není snadné získat nástroje, které mají požadované funkce a přitom nevyžadují speciální znalosti

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Rovnou pro lidi:

- pro lexikografy
  - paralelní konkordance
  - extrakce ekvivalentů slov nebo kolokací
- pro překladatele
  - paralelní konkordance
  - překladová paměť (*Translation Memory*)
  - automatická písařka
    (nabízí nejpravděpodobnější pokračování)
- pro učitele a studenty cizích jazyků
- pro translatology, literární vědce, komparatisty, dialektology
- pro ostatní lingvisty taky!

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

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*)
- zjednoznačňování interpretace textu v jednom jazyce na základě jazyka druhého

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

determined I

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 II

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 III

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 IV

Even before the diagnosis was confirmed, the Odones, 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 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.

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Paralelní korpusy s češtinou

- 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: *22k vět z Wall Street Journal, 53k vět z Reader's Digest*
  `http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.htm`
- Multext/East: 1984 *(George Orwell)* `nl.ijs.si/ME/`
- OPUS: Evropská ústava *(21 jazyků, č.: 11k vět, 128k slov)*, systémová hlášení KDE *(61 jazyků, č.: 90k vět, 367k slov)*, manuály PHP *(22 jazyků, č.: 63k vět, 147k slov)*
  `http://logos.uio.no/opus/`
- 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`
- InterCorp: `https://trnka.ff.cuni.cz/ucnk/intercorp/`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Elektronicky čitelné texty ve více jazycích

- beletrie, zákony EU, www stránky
- Resnik & Smith (2002) The web as a parallel corpus
  `http://www.umiacs.umd.edu/~resnik/pubs.html`
- Baroni, Kilgariff, Pomikálek, Rychlý: BootCat
  `http://corpora.fi.muni.cz/bootcat`

Nebo naskenovat ...

...

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Korpusy prohledávatelné z webového rozhraní

- COMPARA: Portuguese-English
  `http://www.linguateca.pt/COMPARA/Welcome.html`
- Slovene-English Parallel Corpus, asi 1 mil. slov
  `http://nl.ijs.si/elan/`
- Hunglish, Hungarian-English, 54,2 mil. slov
  `http://mokk.bme.hu/resources/hunglishcorpus`
- English-Norwegian Parallel Corpus, obsahuje i španělštinu, němčinu a francouzštinu
  `http://129.177.24.120/webtce.htm`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Různé další odkazy

- Sentence Alignment and Word Alignment: Projects, Papers, Evaluation, etc. `http://www.cs.unt.edu/~rada/wa/`
- Building and Using Parallel Texts: Data Driven Machine Translation and Beyond HLT-NAACL 2003 Workshop, May 31, 2003 `http://www.cs.unt.edu/~rada/wpt/`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Postup přípravy textů pro paralelní korpus

1. akvizice
2. konverze
3. čištění
4. segmentace
5. značkování
6. zarovnávání
7. import do korpusového manažeru

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování znaků

- ISO 8859-2 (ISO Latin 2), CP 1250 (MS Windows), Mac CE, UTF-8 (Unicode)

Kódování formátu

- slova, věty, odstavce, kapitoly; korespondence mezi nimi, pro 2 jazyky:
  - 1 soubor, např. TMX

    `http://www.lisa.org/standards/tmx/`
  - 2 soubory, např. ParaConc, Moore
  - 3 soubory, např. XCES `http://www.xml-ces.org/`

Lingvistické značkování

...

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – vše v jednom souboru
výstup z programu G&C

*** Link: 1 - 1 ***
<Ocs.1.1.2.5> Nemělo smysl zkoušet výtah.
<Oen.1.1.2.5> It was no use trying the lift.
*** Link: 1 - 2 ***
<Ocs.1.1.2.6> I v lepších časech zřídka fungoval a teď se elektrický proud přes den vypínal v rámci úsporných opatření v přípravách na Týden nenávisti.
<Oen.1.1.2.6> Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours.
<Oen.1.1.2.7> It was part of the economy drive in preparation for Hate Week
*** Link: 2 - 1 ***
<Ocs.1.1.2.7> Byt byl v sedmém patře. <Ocs.1.1.2.8> Winston, kterému bylo devětatřicet a měl bércový vřed nad pravým kotníkem, kráčel pomalu a několikrát si cestou odpočinul.
<Oen.1.1.2.8> The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way.

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – vše v jednom souboru
výstup z programu Hunalign hunalign

<P id="cs.1">start</P>	<P id="en.1">start</P>	1.3
<P id="cs.2">ROZHODNUTÍ,</P> — <P id="cs.3">kterým se stanoví den, ke kterému Zásobovací agentura Euratomu přebírá své povinnosti a kterým se schvaluje nařízení Agentury, kterým se stanoví postup při vyrovnání nabídky a poptávky u rud, výchozích materiálů a zvláštních štěpných materiálů</P>	<P id="en.2">DECISION fixing the date on which the Euratom Supply Agency shall take up its duties and approving the Agency Rules of 5 May 1960 determining the manner in which demand is to be balanced against the supply of ores, source materials and special fissile materials</P>	0.035230
<P id="cs.4">KOMISE EVROPSKÉHO SPOLEČENSTVÍ PRO ATOMOVOU ENERGII,</P>	<P id="en.3">THE COMMISSION OF THE EUROPEAN ATOMIC ENERGY COMMUNITY, </P>	0.670313

Kódování formátu – vše v jednom souboru databáze Trados, textový formát I

<ChD>26111999, 10:13:42
<Seg L=DE-DE>Terme werden so eingegeben, wie man sie üblicherweise schreibt.
<Seg L=CS>Výrazy se zadávají v obvyklém formátu.
</TrU>
<TrU>
<ChD>26111999, 10:13:42
<Seg L=DE-DE>Ein- und Ausgabe sind gleichzeitig sichtbar.
<Seg L=CS>Zadané údaje a výsledky jsou viditelné současně.
</TrU>
<TrU>
<ChD>26111999, 10:13:42

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – vše v jednom souboru databáze Trados, textový formát II

<Seg L=DE-DE>Zusammenhänge werden so leichter erkennbar.
<Seg L=CS>Souvislosti tak lépe vyniknou.
</TrU>
<TrU>
<ChD>26111999, 10:13:43
<Seg L=DE-DE>Vorangegangene Eingaben werden gesichert.
<Seg L=CS>Chyba v zadaných údajích je hned patrná.
</TrU>

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – vše v jednom souboru formát TMX I

```
<tu tuid="3589" datatype="Text" changedate="19991126T101342Z">
<tuv lang="DE-DE">
<seg>Terme werden so eingegeben, wie man sie üblicherweise schreibt.</seg>
</tuv>
<tuv lang="CS">
<seg>Výrazy se zadávají v obvyklém formátu.</seg>
</tuv>
</tu>
<tu tuid="3590" datatype="Text" changedate="19991126T101342Z">
<tuv lang="DE-DE">
<seg>Ein- und Ausgabe sind gleichzeitig sichtbar.</seg>
</tuv>
<tuv lang="CS">
<seg>Zadané údaje a výsledky jsou viditelné současně.</seg>
</tuv>
</tu>
<tu tuid="3591" datatype="Text" changedate="19991126T101342Z">
```


Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – vše v jednom souboru formát TMX II

```
<tuv lang="DE-DE">
<seg>Zusammenhänge werden so leichter erkennbar.</seg>
</tuv>
<tuv lang="CS">
<seg>Souvislosti tak lépe vyniknou.</seg>
</tuv>
</tu>
<tu tuid="3592" datatype="Text" changedate="19991126T101343Z">
<tuv lang="DE-DE">
<seg>Vorangegangene Eingaben werden gesichert.</seg>
</tuv>
<tuv lang="CS">
<seg>Chyba v zadaných údajích je hned patrná.</seg>
</tuv>
</tu>
```


Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – dva soubory
výstup z programu ParaConc

...
<seg id="8">Nemělo smysl zkoušet výtah. </seg>
<seg id="9">I v lepších časech zřídka fungoval a teď se elektrický proud přes den vypínal v rámci úsporných opatření v přípravách na Týden nenávisti. </seg>
<seg id="10">Byt byl v sedmém patře. Winston, kterému bylo devětatřicet a měl bércový vřed nad pravým kotníkem, kráčel pomalu a několikrát si cestou odpočinul. </seg>
...
...
<seg id="8">It was no use trying the lift. </seg>
<seg id="9">Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours. It was part of the economy drive in preparation for Hate Week </seg>
<seg id="10">The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way.</seg>
...

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – tři soubory
formát XCES v korpusu OPUS – cs

```
...
<s id="s18.2">
<w id="w18.2.1">Ve</w>
<w id="w18.2.2">svých</w>
<w id="w18.2.3">vztazích</w>
<w id="w18.2.4">s okolním</w>
<w id="w18.2.5">světem</w>
<w id="w18.2.6">Unie</w>
<w id="w18.2.7">zastává</w>
<w id="w18.2.8">a podporuje</w>
<w id="w18.2.9">své</w>
<w id="w18.2.10">hodnoty</w>
<w id="w18.2.11">a zájmy</w>
<w id="w18.2.12">.</w>
</s>
...
```


Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

kódování formátu – tři soubory
formát xces v korpusu opus – en

```
<s id="s18.2">
<chunk id="c18.2-1" type="pp">
<w id="w18.2.1" tree="in" lem="in" pos="in">in</w>
</chunk>
<chunk id="c18.2-2" type="np">
<w id="w18.2.2" tree="pp$" lem="its" pos="prp$">its</w>
<w id="w18.2.3" tree="nns" lem="relation" pos="nns">relations</w>
</chunk>
...
<chunk id="c18.2-7" type="vp">
<w id="w18.2.11" tree="md" lem="shall" pos="md">shall</w>
<w id="w18.2.12" tree="vv" lem="uphold" pos="vb">uphold</w>
<w id="w18.2.13" tree="cc" lem="and" pos="cc">and</w>
<w id="w18.2.14" tree="vv" lem="promote" pos="vb">promote</w>
...
<w id="w18.2.19" tree="sent" lem="." pos=".">.</w>
</s>
```


Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – tři soubory
formát XCES v korpusu OPUS – csen

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE cesAlign PUBLIC "-//CES//DTD XML cesAlign//EN" "">
<cesAlign fromDoc="cs/C2004310CS.01001101.xml"
toDoc="en/C2004310EN.01001101.xml" version="1.0">
<linkGrp targType="s" fromDoc="cs/C2004310CS.01001101.xml"
toDoc="en/C2004310EN.01001101.xml">
<link certainty="0" id="SL0.1" xtargets="s1.1;s1.1" />
<link certainty="13" id="SL1.1" xtargets="s2.1;s2.1" />
...
<link certainty="29" id="SL17.2" xtargets="s18.2;s18.2" />
...
```


Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – tři soubory
výstup ze zarovnávače GMA

1367 <=> 1341
1368 <=> 1342
1369 <=> 1343
1370 <=> 1344
1371 <=> 1345,1346
1372 <=> 1347
1373 <=> 1348,1349
1374 <=> omitted
1375,1376 <=> 1350
1377,1378 <=> 1351
1379 <=> 1352
1380 <=> 1353
1381 <=> 1354
1382 <=> 1355
1383 <=> 1356

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Kódování formátu – tři soubory
výstup ze zarovnávače Hunalign

0	0	1.3
1	1	0.0352308
3	2	0.670313
4	3	2.16048
5	4	0.571795
6	5	0.442454
7	6	0.883784
8	7	1.7875
9	8	0.44718
10	9	1.788
11	10	0.394338
12	11	1.788
13	12	0.525556
14	13	1.39146
15	14	1.788
16	15	0.423446

▸ hunalign

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Použitelné z webového rozhraní

- System Quirk: Text Alignment Server
  `http://www.computing.surrey.ac.uk/SystemQ/align/`
- Corpógrafo, a web-based corpora linguistics tool
  `http://www.linguateca.pt/corpografo/`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Postup přípravy textů pro paralelní korpus

1. akvizice
2. konverze
3. čištění
4. segmentace
5. značkování
6. zarovnávání
7. import do korpusového manažeru

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Nástroje na poloautomatické zarovnávání

– jako součást programového balíku pro podporu překladatele (CAT) - provádí i konverzi a segmentaci, např.:

- Trados - „inteligentní“ zarovnávání, ale $$$
  `http://www.trados.com`
- Déjà Vu 3 - funkční součást demoverze, jen základní funkce `http://www.atril.com`
- CypreSoft TRANS Suite 2000 Align - freeware, základní funkce i párování bez ohledu na pořadí segmentů
  `http://www.cypresoft.com`
- SDLX `http://www.sdlintl.com`
- Star Transit `http://www.star-ag.ch`
- WordFast, makra do MS Wordu `http://www.wordfast.org`
- WordFisher, dtto `http://www.wordfisher.com`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Funkce poloautomatických nástrojů I

Konverze formátů

- pouze text
- textové editory (Word, RTF, OpenOffice, WordPerfect, . . .)
- prezentace (PowerPoint, . . .)
- tabulkové procesory (Excel, . . .)
- databáze (Access, . . .)
- DTP (FrameMaker, PageMaker, QuarkXPress, InDesign, . . .)
- značkované texty (HTML, SGML/XML, TMX, . . .)
- lokalizace softwaru (Interleaf, soubory nápovědy, C, Java, GNU Gettext, . . .)
- formáty CAT (Trados, TMX, . . .)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Funkce poloautomatických nástrojů II

Konverze kódování znaků

- ISO 8859-2 (ISO Latin 2)
- CP 1250 (MS Windows)
- Mac CE
- Unicode (UTF-8, . . .)

Segmentace

- na věty, nadpisy, položky seznamů, popisky obrázků
- podle odstavců (¶) nebo již provedené částečné segmentace
- podle typických zakončení věty: ⟨interpunkce⟩ ⟨mezera⟩
- výjimky: zkratky, čísla

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Funkce poloautomatických nástrojů III

Automatické zarovnávání

- sekvenčně podle segmentů
- podle nadpisů podle formátování
- podle délky segmentů
- podle pravděpodobných ekvivalentů - "anchor points" (čísla, podobné řetězce, překlady slov podle slovníku)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Funkce poloautomatických nástrojů IV

Kontrola a opravy automatického zarovnávání

- paralelní prohlížení
- spojování po sobě jdoucích segmentů
- rozdělování segmentů
- mazání segmentů
- změna pořadí segmentů
- zarovnávání segmentů $1:n, n:1, n:n$
- korespondence křížem

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Nástroje na poloautomatické zarovnávání

– jako součást programového balíku pro jako součást programového balíku pro zpracování paralelních textů, např.:

- Logiterm (Terminotix, Inc.) `http://www.terminotix.com`
- MultiTrans `http://www.multicorpora.com`
- ParaConc `http://www.ruf.rice.edu/~barlow/parac.html`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Nástroje na automatické zarovnávání I

Podle délky segmentů ve znacích

- Gale&Church 1991 – Vanilla Aligner
  `http://www.research.att.com/~kwc/publications.html`, `http://nl.ijs.si/telri/Vanilla/`, `http://www.issco.unige.ch/tools/`, `http://spraakbanken.gu.se/lb/downloads.html`, `evert@IMS.Uni-Stuttgart.DE`
  (EasyAlign - součást IMS CWB)

Podle délky segmentů ve slovech

- Brown et al. 1991

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Nástroje na automatické zarovnávání II

Podle "anchor points"

- distribuce ekvivalentů Kay&Röscheisen 1993
- čísla, formátování, podobné řetězce
- dvoujazyčný slovník Melamed 1996
  `http://www.cs.nyu.edu/~melamed/GMA/docs/README.htm`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Nástroje na automatické zarovnávání III

Kombinace více metod

- Moore 2002
  `http://research.microsoft.com/research/downloads/`
  - předběžné zarovnání podle délky
  - extrakce dvoujazyčného slovníku (stochastickou metodou)
  - přesnější zarovnání podle slovníku
- HunAlign `http://mokk.bme.hu/resources/hunalign`
  - kombinuje zarovnání podle délky, podle ekvivalentů ze slovníku i stochastickou metodu
  - nastavením parametrů lze přizpůsobit konkrétní dvojici jazyků

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Čím se měří úspěšnost zarovnávání I

Pokrytí (recall)

Porovnává se počet správně určených korespondencí (correct links) se skutečným stavem, tedy celkovým počtem korespondencí v souboru (reference links).

$$pokrytí = \frac{počet\ správně\ určených\ korespondencí}{počet\ korespondencí\ v\ souboru}$$

Přesnost (precision)

Porovnává se počet správně určených korespondencí (correct links) s počtem navržených korespondencí ve výsledku zarovnání (test links)

$$přesnost = \frac{počet\ správně\ určených\ korespondencí}{počet\ korespondencí\ ve\ výsledku}$$

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Čím se měří úspěšnost zarovnávání II

Míra F (F-measure)

harmonický průměr pokrytí a přesnosti

$$\textit{míra } F = 2 \times \frac{\textit{pokrytí} \times \textit{přesnost}}{\textit{pokrytí} + \textit{přesnost}}$$

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Ukázky výsledků I

AC – 46+46 documents from the English-Czech part of Acquis Communautaire (roughly 1%); all noise was retained (omissions, results of different segmentation rules); segments = paragraphs

1984 – George Orwell's novel, English and Czech (result of the project Multext-East)

FR7 – Seven French fiction/essay books + Czech translations

Results were compared with hand-corrected alignment of full texts:

Text	Cz words	L2 words	Cz segs	L2 segs	All links	1:1 links
AC	62,010	74,986	3,025	2,699	2,685	89%
1984	99,099	121,661	6,756	6,741	6,657	97%
FR7	289,003	337,226	21,936	21,746	21,207	95%

Ukázky výsledků II

	Ref.	Test	Correct	Recall	Prec.	F-measure
AC						
GC	2700	2683	2225	82.41	82.93	82.67
Mmd$^+$	2700	2686	2492	**92.30**	92.78	**92.54**
Mre	2700	2313	2218	82.15	95.89	88.49
Mre$^+$	2700	2375	2308	85.48	**97.18**	90.96
1984						
GC	6657	6633	6446	**96.83**	97.18	97.01
Mmd$^+$	6657	6606	6287	94.44	95.17	94.81
Mre	6657	6167	6110	91.78	99.08	95.29
Mre*	6657	6370	6320	94.94	99.22	97.03
Mre$^+$	6657	6441	6402	96.17	**99.39**	97.76
Hun	6657	6689	6535	98.17	97.70	**97.93**
F7						
GC	21207	20868	19427	**91.61**	93.09	92.34
Mre	21207	19512	18801	88.65	**96.36**	**92.35**
Mmd	21207	21057	16161	76.21	76.68	76.44

Ukázky výsledků III

	Ref.	Test	Correct	Recall	Prec.	F-measure
AC						
GC	2391	2248	2156	90.17	95.91	92.95
Mmd$^+$	2391	2354	2304	96.36	**97.88**	**97.11**
Mre	2391	2313	2218	92.76	95.89	94.30
Mre$^+$	2391	2375	2308	**96.53**	97.18	96.85
1984						
GC	6440	6438	6274	97.42	97.45	97.44
Mmd$^+$	6404	6301	6287	97.62	**99.78**	98.69
Mre	6440	6167	6110	94.88	99.08	96.93
Mre*	6440	6370	6320	98.14	99.22	98.67
Mre$^+$	6440	6441	6402	**99.41**	99.39	**99.40**
Hun	6440	6479	6386	99.16	98.56	98.86
F7						
GC	20116	19220	19427	92.62	**96.94**	94.73
Mre	20116	19512	18801	**93.46**	96.36	**94.89**
Mmd	20116	19714	15539	77.25	78.82	78.03

Ukázky výsledků IV

Ranking for F-measure (all links)

Rank	AC	1984	F7
1.	92.54 Mmd^{+}	97.93 Hun	92.35 Mre
2.	90.96 Mre^{+}	97.76 Mre^{+}	92.34 GC
3.	88.49 Mre	97.03 Mre*	76.44 Mmd
4.	82.67 GC	97.01 GC	
5.		95.29 Mre	
6.		94.81 Mmd^{+}	

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

S ParaConkem

- Vstup: dva soubory v textovém formátu, kódování Windows nebo UTF-8, s hranicemi odstavců
- Co pomáhá:
  - Zarovnání po odstavcích
  - Označené hranice vět
  - Označené sekce (kapitoly)
  - Zarovnání po větách

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Word&ParaConc à la InterCorp

`http://ucnk.ff.cuni.cz/intercorp/?req=id:5` ▸ ukázky

1. Načtení textu do editoru MS Word
2. „Vyčištění“ textu
3. Oddělení odstavců prázdným řádkem
4. Export z MS Wordu pomocí makra ICorpExport do textového formátu (označení odstavců ... , kódování Windows podle jazyka, např CP1250)
5. Očíslování odstavců ( ), označení vět v českém textu (<s>...</s>), očíslování vět (<s id=...>)
6. Načtení do ParaConku jako „Not Aligned“
7. Oprava odlišného počtu odstavců spojením/rozdělením odstavců v cizím jazyce
8. Oprava zarovnání na věty (nepovinné)
9. Export z ParaConku do dvou souborů se značkami pro segmenty (<seg id=...>...</seg>)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Bolavá místa při přípravě textů

- zarovnání odstavců
  (i při stejném počtu odstavců může dojít k posunutí)
- určení hranic vět
  (není univerzální automatická metoda, která nevyžaduje další znalosti – např. seznamy zkratek)
- zarovnání vět
  (automatická metoda nefunguje na 100%)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Řešení bolavých míst

Řešení v ParaConku

- zarovnání odstavců: ruční spojování/dělení
- určení hranic vět: seznam zkratek, ruční opravy
- zarovnání vět: ruční spojování/dělení

Problémy:

- ParaConc nefunguje na 100%
- hodně ruční práce

Ale: Při troše štěstí a pečlivé ruční práci 100% výsledek

Řešení mimo ParaConc

- využití jiného zarovnávače k zarovnání odstavců
- využití jiného zarovnávače k zarovnání vět

Ale: pak je třeba určit hranice vět ve všech jazycích

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Plán

Zarovnat před načtením do ParaConku

- kdo má Linux, může hned
- kdo nemá, musí ještě chvilku počkat

Zarovnávání on-line

- spouštění zarovnávače z webového rozhraní
- spouštění děliče vět pro daný jazyk z webového rozhraní

Možnosti

- zarovnání odstavců: stačí zarovnávač
- zarovnání vět: je třeba dělič

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Zarovnání odstavců – integrace do postupu InterCorp

1. Načtení textu do editoru MS Word
2. „Vyčištění“ textu
3. Oddělení odstavců prázdným řádkem
4. Export z MS Wordu pomocí makra ICorpExport
5. Očíslování odstavců, označení a očíslování vět v českém textu
6. Zarovnání odstavců v externím zarovnávači
7. Načtení do ParaConku jako „Not Aligned“
8. Oprava odlišného počtu odstavců spojením/rozdělením odstavců v cizím jazyce
9. Oprava zarovnání na věty (nepovinné)
10. Export z ParaConku do dvou souborů se značkami pro segmenty (<seg id=...>...</seg>)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Zarovnání odstavců – integrace do postupu InterCorp

1. Načtení textu do editoru MS Word
2. „Vyčištění“ textu
3. Oddělení odstavců prázdným řádkem
4. Export z MS Wordu pomocí makra ICorpExport
5. Očíslování odstavců, označení a očíslování vět v českém textu
6. Zarovnání odstavců v externím zarovnávači
7. Načtení do ParaConku jako „Not Aligned“
8. Oprava odlišného počtu odstavců spojením/rozdělením odstavců v cizím jazyce
9. Oprava zarovnání na věty (nepovinné)
10. Export z ParaConku do dvou souborů se značkami pro segmenty (<seg id=...>...</seg>)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Zarovnání vět – integrace do postupu InterCorp

1. Načtení textu do editoru MS Word
2. „Vyčištění“ textu
3. Oddělení odstavců prázdným řádkem
4. Export z MS Wordu pomocí makra ICorpExport
5. Očíslování odstavců, označení a očíslování vět v českém textu
6. Označení vět v cizím textu v externím děliči vět
7. Zarovnání vět v externím zarovnávači
8. Načtení do ParaConku jako „Not Aligned“
9. Načtení do ParaConku jako „Aligned“
10. Oprava odlišného počtu odstavců spojením/rozdělením odstavců v cizím jazyce
11. Oprava zarovnání na věty (nepovinné)
12. Export z ParaConku do dvou souborů se značkami pro segmenty (<seg id=...>...</seg>)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Zarovnání vět – integrace do postupu InterCorp

1. Načtení textu do editoru MS Word
2. „Vyčištění“ textu
3. Oddělení odstavců prázdným řádkem
4. Export z MS Wordu pomocí makra ICorpExport
5. Očíslování odstavců, označení a očíslování vět v českém textu
6. Označení vět v cizím textu v externím děliči vět
7. Zarovnání vět v externím zarovnávači
8. Načtení do ParaConku jako „Not Aligned“
9. Načtení do ParaConku jako „Aligned“
10. Oprava odlišného počtu odstavců spojením/rozdělením odstavců v cizím jazyce
11. Oprava zarovnání na věty (nepovinné)
12. Export z ParaConku do dvou souborů se značkami pro segmenty (<seg id=...>...</seg>)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Děliče vět: *Sentence splitters, Segmenters, Tokenizers, Sentencers*

- tokenizér/segmentátor Pavla Květoně pro češtinu, používá se v projektu InterCorp, další aplikace třeba dohodnout s autorem
- MULTEXT/MULTEXT-East
  `http://nl.ijs.si/ME/CD/docs/mte-tools.html` – segmenter v sadě nástrojů ke zpracování bulharštiny, češtiny, angličtiny, estonštiny, maďarštiny, rumunštiny, slovinštiny, francouzštiny, španělština, nizozemštiny, němčiny, italštiny
- UNIVERSITY OF ILLINOIS Sentence Segmentation tool
  `http://l2r.cs.uiuc.edu/~cogcomp/atool.php?tkey=SS` volně pro akademické účely, zdrojový kód lze upravovat, perl, angličtina, seznam titulů
- Segmentátor pro angličtinu a hebrejštinu jako modul perlu, lze upravovat `http://search.cpan.org/~shlomoy/`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Zarovnávač: *Hunalign*

- `http://mokk.bme.hu/resources/hunalign`
- vstup: dva segmentované soubory, segmenty odděleny novým řádkem
- výstup: soubor se třemi sloupci ▸ text
  nebo jen s pořadovými čísly segmentů ▸ čísla
- dostane-li slovník ▸ slovník, kombinuje lexikální informace s metodou Gale-Church
- nemá-li slovník, vytvoří si ho v prvním kroku sám z korespondencí podle metody Gale-Church, a podle slovníku pak v druhém kroku zarovnání zpřesní
- nedokáže vytvářet korespondence křížem

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Hunalign – další funkce

- u každé korespondence je hodnocení spolehlivosti
- výstupní filtry:
  - jen korespondence 1:1
  - jen korespondence, před nimž a za nimiž jsou korespodence 1:1
  - potlačit korespondence s hodnocením nižším než zadaná hodnota
  - …
- výpočet přesnosti a pokrytí vzhledem ke vzoru

Jak zlepšit výsledek? Slovník, lematizace vstupů.

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Korpusové manažery

- ParaConc `http://www.ruf.rice.edu/~barlow/parac.html`
- Uplug `http://stp.ling.uu.se/~joerg/uplug/`
- COMPARA `http://www.linguateca.pt/COMPARA/Welcome.html`, IMS CWB `http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/`
- MultiLingual Concordancer in Java `http://www.lancs.ac.uk/staff/piaosl/research/download/download.htm`

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Obvyklé vyhledávací funkce

- dotaz na libovolný jazyk nebo více jazyků zároveň (paralelní hledání)
- zadání dotazu regulárním výrazem
- hledání podle značek
- omezení prohledávaných textů:
  - bibliografické údaje
  - originál nebo překlad
  - jazyková varianta (britská/americká angličtina)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Zobrazení výsledku dotazu

- kontext: segment nebo KWiC
- zadání/zjištění ekvivalentů, BiKWiC
- třídění podle KW, kontextu
- zobrazení/potlačení značek
- zobrazení kolokací
- údaje o zarovnání (n:n, spolehlivost)
- poznámky překladatele

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

statistiky

- frekvence tvarů
- kolokace
- frekvence kolokací
- distribuce forem
- distribuce zdrojů

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních korpusech

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

- Překlad s využitím paralelního korpusu (překladová paměť, překlad podle příkladů, statistický překlad)
- Extrakce dvoujazyčného slovníku (zarovnávání slov, víceslovných výrazů)

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Obsah

Úvod
- O semináři ...
- Korpusy a paralelní korpusy
- K čemu je paralelní korpus?
- Ukázky paralelních konkordancí

Existující korpusy a zdroje dat
- Kde je něco česky?
- Další paralelní korpusy

Technické aspekty
- Formát dat
- Programové nástroje

Příprava textů
- Poloautomatické zarovnávání
- Automatické zarovnávání
- Hodnocení výsledků zarovnávání
- Recept na (paralelní) korpus

Hledání v paralelních korpusech

Další využití paralelních korpusů

Přílohy

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních

Původní text

"Shoo!" said Mr. Dursley loudly.The cat didn’t move. It just gave him a stern look.Was this normal cat behavior? Mr. Dursley wondered.Trying to pull himself together, he let himself into the house.He was still determined not to mention anything to his wife.

"Všššc!" sykl pan Dursley nahlas.

Kočka se ani nepohnula, jenom se na něj přísně podívala.Pan Dursley chvilku uvažoval, jestli se kočky takhle chovají normálně.Zatímco se nutil ke klidu, otevřel si domovní dveře;ještě pořád nehodlal manželce nic říkat.

▸ postup

Po exportu z Wordu (.txt)

"Shoo!" said Mr. Dursley loudly.The cat didn’t move. It just gave him a stern look.Was this normal cat behavior? Mr. Dursley wondered.Trying to pull himself together, he let himself into the house.He was still determined not to mention anything to his wife.

"Všššc!" sykl pan Dursley nahlas.

 Kočka se ani nepohnula, jenom se na něj přísně podívala.Pan Dursley chvilku uvažoval, jestli se kočky takhle chovají normálně.Zatímco se nutil ke klidu, otevřel si domovní dveře;ještě pořád nehodlal manželce nic říkat.

▸ postup

Po očíslování odstavců (.txt1)

"Shoo!" said Mr. Dursley loudly.The cat didn’t move. It just gave him a stern look.Was this normal cat behavior? Mr. Dursley wondered.Trying to pull himself together, he let himself into the house.He was still determined not to mention anything to his wife.

"Všššc!" sykl pan Dursley nahlas.

 Kočka se ani nepohnula, jenom se na něj přísně podívala.Pan Dursley chvilku uvažoval, jestli se kočky takhle chovají normálně.Zatímco se nutil ke klidu, otevřel si domovní dveře;ještě pořád nehodlal manželce nic říkat.

▸ postup

Po označení českých vět (.txt1)

"Shoo!" said Mr. Dursley loudly.The cat didn’t move. It just gave him a stern look.Was this normal cat behavior? Mr. Dursley wondered.Trying to pull himself together, he let himself into the house.He was still determined not to mention anything to his wife.

<s id="23.1">"Všššc!" sykl pan Dursley nahlas.</s>

 <s id="24.1">Kočka se ani nepohnula, jenom se na něj přísně podívala.</s> <s id="24.2">Pan Dursley chvilku uvažoval, jestli se kočky takhle chovají normálně.</s> <s id="24.3">Zatímco se nutil ke klidu, otevřel si domovní dveře;</s> <s id="24.4">ještě pořád nehodlal manželce nic říkat.</s>

postup

Po zarovnání (.seg)

<seg id="89">"Shoo!" said Mr. Dursley loudly.</seg> <seg id="90">The cat didn’t move. It just gave him a stern look.</seg> <seg id="91">Was this normal cat behavior? Mr. Dursley wondered.</seg> <seg id="92">Trying to pull himself together, he let himself into the house.</seg> <seg id="93">He was still determined not to mention anything to his wife.</seg>

<s id="23.1"><seg id="89">"Všššc!" sykl pan Dursley nahlas.</seg></s>

 <s id="24.1"><seg id="90">Kočka se ani nepohnula, jenom se na něj přísně podívala.</seg></s> <s id="24.2"><seg id="91">Pan Dursley chvilku uvažoval, jestli se kočky takhle chovají normálně.</seg></s> <s id="24.3"><seg id="92">Zatímco se nutil ke klidu, otevřel si domovní dveře;</seg></s> <s id="24.4"><seg id="93">ještě pořád nehodlal manželce nic říkat.</seg></s>

Slovník pro *Hunalign*

▸ hunalign

průkopnický @ innovative
průkopnický @ pioneering
průkopníci @ pioneers
průkopník @ pathfinder
průkopník @ pioneer
průkopník @ spearhead
průkopník @ trailblazer
průlet @ fly-by
průlez @ hatchway
průlez @ manhole
průliv @ channel
průliv @ kyle
průlom @ breach
průlom @ breakout
průlom @ breakthrough
průlom @ rupture
průlomy @ breakthroughs

Úvod
O semináři ...
Korpusy a paralelní korpusy
K čemu je paralelní korpus?
Ukázky paralelních konkordancí
Existující korpusy a zdroje dat
Kde je něco česky?
Další paralelní korpusy
Technické aspekty
Formát dat
Programové nástroje
Příprava textů
Poloautomatické zarovnávání
Automatické zarovnávání
Hodnocení výsledků zarovnávání
Recept na (paralelní) korpus
Hledání v paralelních