

Paralelní korpusy

seminář ÚČNK, 2. dubna 2009

Alexandr Rosen

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

1 Úvod

2 Existující korpusy a zdroje dat

3 Technické aspekty

4 Příprava textů

5 Hledání v paralelních korpusech

6 Další využití paralelních korpusů

7 Různé

8 Web jako paralelní korpus

9 Přílohy

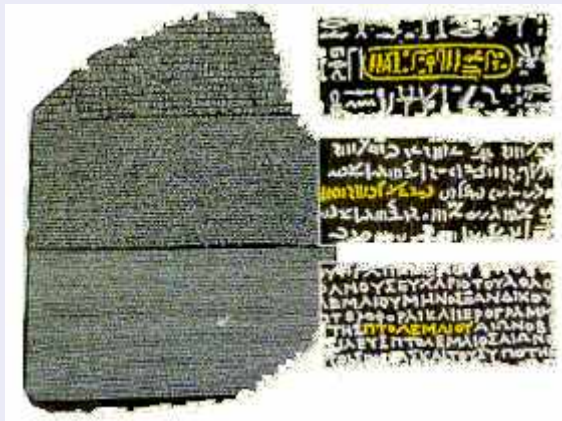
Osnova

- 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: počítačnická lexikografie, hledání v cizojazyčných textech, strojový nebo počítačem podporovaný překlad, ...

- **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*.

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 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 lexikografy
 - ▶ paralelní konkordance
 - ▶ extrakce ekvivalentů slov nebo kolokací
- pro překladatele
 - ▶ paralelní konkordance
 - ▶ překladová paměť (*Translation Memory*)
 - ▶ automatická písárka
(nabízí nejpravděpodobnější pokračování)
- pro učitele a studenty cizích jazyků
- 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*)
- zjednodušování interpretace textu v jednom jazyce
na základě jazyka druhého

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

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 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 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 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 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 holčkám, 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 Existující korpusy a zdroje dat**
- 3 Technické aspekty
- 4 Příprava textů
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů
- 7 Různé
- 8 Web jako paralelní korpus
- 9 Přílohy

- Kde je něco česky?
- Další paralelní korpusy

Paralelní korpusy s češtinou

- 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: *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/>

Paralelní korpusy s češtinou – pokr.

- **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: Acquis, EU, Navajo, Gnome, KDE, e-books (č.: 1,4 mil. vět, 21 mil. slov)**
<http://ufal.mff.cuni.cz/czeng/>
- **InterCorp: <http://korpus.cz/corpora/intercorp/>**

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: WebBootCat – nástroj na generování korpusu podle zadaných klíčových slov
<http://corpora.fi.muni.cz/bootcat>

Nebo naskenovat ...

...

- Kde je něco česky?
- Další paralelní korpusy

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>

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/>

- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty**
- 4 Příprava textů
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů
- 7 Různé
- 8 Web jako paralelní korpus
- 9 Přílohy

- Formát dat
- Programové nástroje

Postup přípravy textů pro paralelní korpus

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

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

...

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.

Kódování formátu – vše v jednom souboru

výstup z programu Hunalign ▶ hunalign

<P id="cs.1">start</P>

<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="cs.4">KOMISE EVROPSKÉHO SPOLEČENSTVÍ PRO ATOMOVOU ENERGII,</P>

<P id="en.1">start</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>

<P id="en.3">THE COMMISSION OF THE EUROPEAN ATOMIC ENERGY COMMUNITY,</P>

1.3

0.035230

0.670313

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

<TrU>
<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

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>

Kódování formátu – 1 soubor, 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>
```

Kódování formátu – 1 soubor, formát TMX II

```
<tu tuid="3591" datatype="Text" changedate="19991126T101342Z">  
<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>
```

Kódování formátu – 2 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">`Byl by 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>`

Kódování formátu – 3 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>
```

```
...
```

Kódování formátu – 3 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>
```


Kódování formátu – 3 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" />
...
```

Kódování formátu – 3 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

Kódování formátu – 3 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

- Formát dat
- Programové nástroje

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/>

- **Segmentace a zarovnání:**

<http://chomsky.ruk.cuni.cz/hunalign.>

Napište si vyučujícímu o login a heslo.

- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty
- 4 Příprava textů**
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů
- 7 Různé
- 8 Web jako paralelní korpus
- 9 Přílohy

Postup přípravy textů pro paralelní korpus

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

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

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>

File alignment

File Edit Tools

Source	Target
And the earth was without form, and void;	Země byla pustá a prázdná
and darkness [was] upon the face of the deep.	a nad propastnou tůňí byla tma.
And the Spirit of God moved upon the face of the waters.	Ale nad vodami vznášel se duch Boží.
And God said,	I řekl Bůh:
Let there be light:	"Buď světlo!"
and there was light.	A bylo světlo.
And God saw the light that [it was] good:	Viděl, že světlo je dobré, a oddělil světlo od tmy.
and God divided the light from the darkness.	Světlo nazval Bůh dnem a tmou nazval nocí.
And God called the light Day, and the darkness he called	Byl večer a bylo jitro, den první.
And the evening and the morning were the first day.	Na počátku bylo Slovo, to Slovo bylo u Boha, to Slovo
In the beginning was the Word, and the Word was with Go	To bylo na počátku u Boha.
The same was in the beginning with God.	Všechno povstalo skrze ně a bez něho nepovstalo nic
All things were made by him:	V něm byl život a život byl světlo lidí
And God saw the light that [it was] good:	Viděl, že světlo je dobré, a oddělil světlo od tmy.

Join source Split source Erase source

Join target Split target Erase target

TRANS Suite 2000 Align - bible [bible_en.doc -> bible_cs.doc]

File Edit Source Block Target Block Project Jump Tools View Help

Arial 20 Arial

Whole word only Match Case

In the beginning God created the heaven and the earth.	Na počátku stvořil Bůh nebe a zemi.
And the earth was without form, and void;	Země byla pustá a prázdná a nad propastnou tůňí byla tma.
and darkness [was] upon the face of the deep.	Ale nad vodami vznášel se duch Boží.
And the Spirit of God moved upon the face of the waters.	I řekl Bůh:
And God said, Let there be light:	"Bud' světlo!"
and there was light.	A bylo světlo.

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, ...)
- značkové texty (HTML, SGML/XML, TMX, ...)
- lokalizace softwaru (Interleaf, soubory nápovědy, C, Java, GNU Gettext, ...)
- formáty CAT (Trados, TMX, ...)

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

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)

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

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>

ParaConc - [Alignment English (United Kingdom) - Czech (ENCAT.TXT - CSCAT.txt): Segments]	
File Alignment Search Frequency Window Info	
By 1986 , more than 20 research groups had found H. pylori in the stomachs of patients with gastritis.	Do roku 1989 objevilo více než dvacet vědeckých týmů v žaludcích svých pacientů bakterie rodu H. pylori. To podnítilo další výzkumy.
Studies that followed found that adding a two-week dose of generic antibiotics and bismuth to traditional ulcer medications can obliterate the bacteria and keep ulcers away in at least 75 percent of cases.	Všechny potvrdily Marshallovy základní poznatky v tom smyslu, že podáváním běžných antibiotik a v některých případech i bizmutu v kombinaci s tradičními protivředovými přípravky po dobu dvou až čtyř týdnů lze nejméně u pětasedmdesáti procent případů docílit zničení bakterií a prevence vředového onemocnění žaludku a dvanáctníku.
Two more studies involving almost 135,000 people demonstrated the life-saving possibilities of eradicating the bacteria.	Lékaři se dlouho domnívali, že nedostatečnost žaludečního epitelu je dědičná porucha. Ale vědci došli k závěru, že sedmdesát procent osob, infikovaných bakteriemi rodu H. pylori, předává choroboplodné zárodky svým partnerům, a pouze čtyřicet procent svým potomkům. To ukazuje, že hlavní příčinou vředového onemocnění je nákaza, nikoli dědičnost. Dvě vědecké studie, vycházející z vyšetření 136 000 osob, ukazují na možnosti záchrany lidských životů.
Researchers found that infection with H. pylori is strongly associated with the development of stomach cancer -- the second-most-common cancer.	Bylo totiž zjištěno, že osoby nakažené bakteriemi H. pylori jsou šestkrát více ohroženy rakovinou žaludku, druhým nejčastějším nádorovým onemocněním na světě.

1 parallel file loaded

839 405 / 731 145

12:58

ParaConc

File Alignment Search Frequency Window Info

Alignment English (United Kingdom) - Czech (bible_win_en.txt - bible_win_cs.txt): Segments

In the beginning God created the heaven and the earth.	Na počátku stvořil Bůh nebe a zemi.
And the earth was without form, and void; and darkness [was] upon the face of the deep.	Země byla pustá a prázdná a nad propastnou tůňí byla tma.
And the Spirit of God moved upon the face of the waters.	Ale nad vodami vznášel se duch Boží.
And God said, Let there be light: and there was light.	řekl Bůh: "Buď světlo! " A bylo světlo.
And God saw the light, that [it was] good: and God divided the light from the darkness.	Viděl, že světlo je dobré, a oddělil světlo od tmy.
And God called the light Day, and the darkness he called Night.	Světlo nazval Bůh dnem a tmou nazval nocí.
And the evening and the morning were the first day.	Byl večer a bylo jitro, den první.
In the beginning was the Word, and the Word was with God, and the Word was God.	Na počátku bylo Slovo, to Slovo bylo u Boha, to Slovo byl Bůh.
The same was in the beginning with God.	To bylo na počátku u Boha.
All things were made by him; and without him was not any thing made that was made.	Všechno povstalo skrze ně a bez něho nepovstalo nic, co jest.
In him was life; and the life was the light of men.	V něm byl život a život byl světlo lidí.

1 parallel file loaded

155 / 106

12:33

ParaConc

File Alignment Search Frequency Window Info

Alignment: English (United Kingdom) - Czech (ENCAT.TXT - CSCAT.txt): Segments

Each spring and summer brought almost manic highs. She had boundless energy that enabled her to go with little sleep for days and effortlessly shed the excess weight she'd gained.	Každé jaro a léto jí zase přineslo radikální zlepšení nálady: měla obrovskou energii, po celé dny jí stačilo minimum spánku a bez námahy shodila veškerou nadbytečnou váhu získanou v zimě.
By her mid-40s, these seasonal ups and downs began impairing her work as a writer.	Okolo pětácti let se toto kolísání nálad podle ročního období začalo záporně projevovat v její spisovatelské práci.
Finally, urged on by her husband, Barry sought professional help.	Na manželovo naléhání nakonec vyhledala lékařskou pomoc.
In 1987 she joined an experimental program at the U.S. National Institute of Mental Health (NIMH) in Bethesda, Md. Research psychiatrist Norman Rosenthal, the program's director, knew just what Barry was going through.	V roce 1987 se stala pacientkou experimentálního programu Národního ústavu pro duševní zdraví (NIMH) v Bethesda, ve státě Maryland. Ředitel programu, psychiatr Norman Rosenthal, velice dobře věděl, co se s paní Barryovou děje.
The gloom of winter had clouded his own moods after he moved from sunny South Africa in 1976.	I jeho náladu zkazily zimní chmury, když se v roce 1976 přistěhoval ze slunečné jižní Afriky.
He also knew he wasn't unique.	Věděl také, že to není nic výjimečného.
"Greek and Roman physicians in ancient	"Už v antických dobách věděli řečtí a římskí

1 parallel file loaded

839 405 / 731 145 12:47

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

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

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>

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

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

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).

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

Přesnost (precision)

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

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

Čím se měří úspěšnost zarovnávání II

Míra F (F-measure)

harmonický průměr pokrytí a přesnosti

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

Ukázky výsledků I

- AC** – 46+46 dokumentů z anglicko-české části **Acquis Communautaire** (asi 1%); se zachováním všech chyb (vynechávky, chybná segmentace); segmenty = odstavce
- 1984** – román **George Orwella**, anglicky a česky (výsledek projektu Multext-East)
- FR7** – sedm **francouzských** monografií (beletrie a literatura faktu) + české překlady

Výsledky byly porovnány s ručně opraveným zarovnáním:

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

Pořadí podle F-measure (všechny korespondence)

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 ⁺	

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

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

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 `<p>...</p>`, kódování Windows podle jazyka, např CP1250)
- 5 Očíslování odstavců (`<p id=...>`), 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>`)

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 zkratk)
- zarovnání vět
(automatická metoda nefunguje na 100%)

Řešení bolavých míst

Řešení v ParaConku

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

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

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/>

Děliče vět II

- **Punkt** (Kiss & Strunk, CL 32 (2006)), implementace v Pythonu <http://nltk.sourceforge.net/> – program se učí zkratky z textu, s jejich pomocí a na základě různých heuristik se pokouší dělit věty

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

Hunalign – další funkce

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

Jak zlepšit výsledek? Slovník, lematizace vstupů.

- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty
- 4 Příprava textů
- 5 Hledání v paralelních korpusech**
- 6 Další využití paralelních korpusů
- 7 Různé
- 8 Web jako paralelní korpus
- 9 Přílohy

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>

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)

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

statistiky

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

Advanced Search ✕

Language: English (United Kingdom) ▼

Enter pattern to search for:

beer ▼

e.g. "colou?r", "\ba\W\w+\Wof\b", "{\b[A-Z]\w*\W}{2}"

<p>Search Syntax</p> <p><input type="radio"/> Text Search</p> <p><input checked="" type="radio"/> Regular Expression</p> <p><input type="radio"/> Tag Search</p>	<p>General Search Control</p> <p><input checked="" type="checkbox"/> Ignore case of letters</p> <p><input type="checkbox"/> Use skipping and equal characters</p> <p><input type="checkbox"/> Sentence mode</p>
--	---

Additional Search Control

Headings/Contexts Edit...

Append search

Options... OK Cancel

ParaConc - [Parallel Concordance - [beer]]

File Search Frequency Display Sgrt Window Info

potato salad and some chicken and **drink a beer and** stay out. Then when winter came, we ...
 ... on his belly. It always won him **a beer**. Now he would do the same -- but ...
 ore flavonoids than light or caramel-colored **beer**, produced **a** drop in platelet clumping simila
 punt of lycopene the body can absorb. **Dark Beer**. You may have heard that men in France, ...
 search suggests that people who **drink dark beer** may reap a similar benefit. "It's not the ...
 ... that one or two daily glasses of **dark beer** or any other alcoholic beverage rich in flavor
 ydy by Folts's researchers showed that **dark beer**, which has more flavonoids than light or cara
 e gang of his buddies laughing and **drinking beer and** setting up a poker game for the ...
rove around eastern Ohio together, drinking beer, shooting up road signs and committing pett
 anded. The two detectives, clutching empty **beer** cans as though they were a couple of ...
 ... He was listening and he'd had a few **beers**, so he might have punched me if ...
 ... better check on her." My colleague **John Beery** took his gear and joined the other ...

... bramborový salát, upekli kuře a šli si s **pivem** sednout ven. Když přišla zima, scházeli jsm
 .. mu postavit na břicho. Vždycky se sázel o **pivo**, že **to** dokáže. Takže teď udělá totéž, ale ...
 ů. Foltsův tým zjistil, že **konzumace černého piva, které má** více flavonoidů než světlé nebo ka
 eré je naše tělo schopno absorbovat. **Černé pivo**. Francouzi, kteří rádi pijí k jídlu červené víno,
 odobně blahodárné účinky může mít i **černé pivo**. "Není **to** v samotném alkoholu," vysvětluje J
 pijí denně jednu nebo dvě sklenice **černého piva** -- nebo jakéhokoli alkoholu bohatého na flav
 ů. Foltsův tým zjistil, že **konzumace černého piva, které má** více flavonoidů než světlé nebo ka
 o, jak venku na parkovišti popíjí s kamarády **pivo a** živě plánuje partičku pokeru, kterou si zah
 tech s Dillonem často jezdili po kraji, popíjeli **pivo**, trefovali se puškou do silničních značek a p
 li kolem něho. Třímali prázdné plechovky od **piva**, jako by se v rozjařené náladě vraceli z ...
 ... s vámi mluvil. Měl už v sobě pár **piv**, a kdybych řekl něco o něm, asi by ...

30 matches English (United Kingdom) - 1st left, 2nd left Strings matching: beer

English (United Kingdom) - C:\Documents and Settings\All Users\Dokumenty\rosen\Parallels\Texts\ReadersDigest\ENCAT.TXT, line: 839 405 / 731 145 13:42

Advanced Search ✕

Language: Czech ▼

Enter pattern to search for:

`\bpravd[a-z]*\b` ▼

e.g. "colou?r", "\ba\W\w+\Wof\b", "{\b[A-Z]\w*\W}{2}"

Search Syntax

Text Search
 Regular Expression
 Tag Search

General Search Control

Ignore case of letters
 Use skipping and equal characters
 Sentence mode

Additional Search Control

Headings/Contexts Edit...
 Append search

Options... OK Cancel

ParaConc - [Parallel Concordance - [\bpravd[a-z]^b]]

File Search Frequency Display Sort Window Info

.. nou bude kurýrovat tygry". Měla úplnou pravdu. Moc rád bych jí vyprávěl tento příběh ...
 ... il jsem. "Nikdo tu není věčně." "Máte pravdu, mladý muži." Začal vyprávět, jak se s ...
 ... vyvrátit mé závěry, asi zjistí, že mám pravdu," prohlásil. Jeden lékař se do toho př ...
 ... příslušností ke společné doktríně. Je pravda, že naše společnost je stratifikovaná, ...
 .. u Kapustičku. Mám všecko!" To dítě má pravdu. Podobenství pro dnešní den Šíleně smu ...
 ... má tam v hloubce nějaký vliv." King má pravdu. Jak ukazují nové výzkumy, sny jsou ča ...
 ... ste, prosím, carltonky." Lauria má ale pravdu v tom, že většina inzerátů na lehké cig ...
 ... věra bortí a do srdcí se vkrádá chlad. Pravda je důležitá. Žena, která slyší příliš ...
 ... e svými hračkami ve vaně? Ať už je ale pravda jakákoli, zdá se, že jsme zbořili barié ...
 ... do klubička a raději promokněte. Nemí pravda, že blesk nikdy neudeří do stejného mís ...
 ... , které o svém těle "víme", prostě není pravda. Neméně úžasné jsou záhady těla, které ..
 ... o doopravdy nemyslel." "Jenže to není pravda." "Ve chvíli, kdy se to děje, to člov ...
 ... knihu nepíše jen jeden člověk." "Je pravda, co se v ní píše?" "Co se tam nepíše ...
 ... that child's going to treat tigers." How right she was. I would have loved to have ...
 ... " I said. "Nothing lasts forever." "How right you are, young man." He told me ...
 to prove me wrong, they might discover I'm right," he told them. One doctor who picked ...
 .. by adherence to a common doctrine. It is true that our society is stratified, and very rigidly
 ... doll. I have everything!" The kid is right. The Terribly, Tragically Sad Man A parable
 ... sort of influence down there." King is right. Dreams, new research is showing, are offer
 .. smoke please try Carltons." But Lauria is right that most low-tar ads just prattle about deliv
 earts growing colder. The healing oxygen is truth. A woman who is hearing too many lies ...
 University, suggests. Regardless of who is right, we seem to have crossed a frontier with ...
 n yourself to getting wet. Incidentally, it isn't true that lightning never strikes twice in the same
 e "know" about our bodies that simply isn't true. And equally amazing are the mysteries abo
 ... and didn't really mean it." "But that isn't true." "At the time when it happens you do ...

Start | NFS Server | Microsoft P... | Technical d... | Vyhledat G... | C:\Docume... | C:\Docume... | ParaConc | 14:30

Hot Words - English (United Kingdom)

Choose hot words to highlight:

Rank	Word
50,23	truth
22,73	true
21,02	right
9,60	ministry
3,43	telling
3,28	lies
2,37	truths
2,09	absolute
1,95	lie
1,83	minitrue

Options...

OK

Cancel



COMPARA complex search - Microsoft Internet Explorer

Soubor Úpravy Zobrazit Oblíbené Nástroje Nápověda

← Zpět → Hledat Oblíbené Média

Adresa <http://www.linguateca.pt/COMPARA/ComplexSearch.html> Přejít Odkazy »

COMPARA complex search

A complex search enables you to carry out more sophisticated queries and choose which parts of COMPARA you wish to use. You can also select different outputs. To do this, follow steps 1 to 4 below.

1. Select language direction

[Help](#)

From Portuguese to English

From English to Portuguese

2. Enter query

Type in search word or expression [Help](#) Type in alignment constraint (optional) [Help](#)

Other searchable features

[Help](#)

- Translators' notes
- Titles
- Foreign words and expressions
- Within-sentence emphasis
- Sentences added to translation
- Sentences deleted from translation
- Sentences reordered in translation
- Sentences joined together in translation
- Sentences split in translation
- All of the above sentence changes

Internet

COMPARA complex search - Microsoft Internet Explorer

Soubor Úpravy Zobrazit Oblíbené Nástroje Nápověda

↶ Zpět ↷ ↻ ↵ Hledat Oblíbené Média

Adresa <http://www.linguateca.pt/COMPARA/ComplexSearch.html> Přejít Odkazy »

3. Do you wish to use the whole corpus?

Yes - Go straight to [step 4](#)
No - Select any of the preferences below

3.1 Restrict language varieties to:

[Help](#)

Portuguese	English
<input type="checkbox"/> Angola	<input type="checkbox"/> South Africa
<input type="checkbox"/> Brazil	<input type="checkbox"/> United Kingdom
<input type="checkbox"/> Mozambique	<input type="checkbox"/> United States
<input type="checkbox"/> Portugal	

3.2 Restrict dates of publication to:

[Help](#)

Source texts first published after

Translations first published after

3.3 Searches to go only from: [Help](#)

source texts to translations
 translations back to source texts

Internet

COMPARA complex search - Microsoft Internet Explorer

Soubor Úpravy Zobrazit Oblíbené Nástroje Nápověda

← Zpět → Hledat Oblíbené Média

Adresa <http://www.linguateca.pt/COMPARA/ComplexSearch.html> Přejít Odkazy »

3.4 Use only the following texts:

[Help](#)

<input type="checkbox"/> EBDL1T1	<input type="checkbox"/> EBDL1T2	<input type="checkbox"/> EBDL2	<input type="checkbox"/> EBDL3T1
<input type="checkbox"/> EBDL3T2	<input type="checkbox"/> EBJB1	<input type="checkbox"/> EBJT1	<input type="checkbox"/> EBJT2
<input type="checkbox"/> EBOW1	<input type="checkbox"/> ESNG1	<input type="checkbox"/> EUHJ1	<input type="checkbox"/> PBAA1
<input type="checkbox"/> PBAD1	<input type="checkbox"/> PBJA1T1	<input type="checkbox"/> PBJA1T2	<input type="checkbox"/> PBMA1
<input type="checkbox"/> PBMA2	<input type="checkbox"/> PBMA3	<input type="checkbox"/> PBPC1	<input type="checkbox"/> PBPM1
<input type="checkbox"/> PBRF1	<input type="checkbox"/> PMMC1	<input type="checkbox"/> PMMC2	<input type="checkbox"/> PPCC1
<input type="checkbox"/> PPCP1	<input type="checkbox"/> PPEQ1	<input type="checkbox"/> PPEQ2	<input type="checkbox"/> PPMC1
<input type="checkbox"/> PPSC1			

4. Choose output

[Help](#)

Concordance Show alignment properties Hide translators' notes

Distribution of forms

Distribution of sources

Combined distribution of Portuguese and English search expressions

Internet

COMPARA - Results for ghost.*: - Microsoft Internet Explorer

Soubor Úpravy Zobrazit Oblíbené Nástroje Nápověda

↩ Zpět → → Hledat Oblíbené Média

Adresa <http://www.linguateca.pt/cgi-bin/paralelo.pl> Přejít Odkazy

Initial search string: **ghost.***. Request for : **distribution of forms, distribution of sources, KWIC Concordance, combined distribution of Portuguese and English search expressions**. Corpus: **From English to Portuguese**.

Results of your search

You are welcome to use the results of your COMPARA search queries for research and education, provided the source is acknowledged. To cite specific texts from the corpus, click on the blue code next to them for their full reference. To cite the corpus as a whole, please refer to: **COMPARA 3.5** <http://www.linguateca.pt/COMPARA/> [13-Jun-2003]

15 instances found.

Distribution of forms

ghost 7
ghostfly 4
ghosts 4

Distribution of sources

PBRF1 3 [32320]
PPCP1 2 [14302]
PBAD1 2 [21042]
EBDL3T2 2 [27683]
EBDL3T1 2 [27683]

Hotovo Internet

COMPARA - Results for ghost.*: - Microsoft Internet Explorer

Soubor Úpravy Zobrazit Obilbené Nástroje Nápočyda

↵ Zpět → → → → Hledat Obilbené Média ↻ ↵ ↵ ↵ ↵ ↵

Adresa <http://www.linguateca.pt/cgi-bin/paralelo.pl> ↻ Přejít Odkazy »

Concordance

Search string: "ghost.*".

EBDL1T1 (1473): 1-1	My fears and anxieties shrank and receded and disappeared, like gibbering ghosts in the light of day.	Os meus medos e ansiedades encolheram, esbateram-se até desaparecerem, como fantasmas balbuciantes à luz do dia.
EBDL1T2 (1473): 1-1	My fears and anxieties shrank and receded and disappeared, like gibbering ghosts in the light of day.	Minhas fobias e ansiedades se encolheram e desapareceram, como fantasmas à luz do dia.
EBDL3T1 (115): 1-1	He visualized a critical work of totally revolutionary form, a concise, comprehensive survey of English literature consisting entirely of questions, elegantly printed with acres of white paper between them, questions that would be miracles of condensation, eloquence and thoughtfulness, questions to read and re-read, questions to brood over, as pregnant and enigmatic as <i>haikus</i> , as memorable as proverbs; questions that would, so to speak, contain within themselves the ghostly , subtly suggested embryos of their own answers.	Visualizou um trabalho crítico de forma totalmente revolucionária, um estudo conciso e alargado sobre literatura inglesa, que consistia inteiramente em questionários elegantemente impressos, com quilómetros de papel em branco pelo meio, questionários que seriam milagres de condensação, eloquência e precisão, questionários para ler e reler e ruminar, tão prenhes e enigmáticos como <i>haikus</i> , tão memoráveis como provérbios; questionários que, por assim dizer, conteriam em si próprios os embriões fantasmática e subtilmente sugeridos das suas próprias respostas.
EBDL3T1 (647): 1-1	In Faustian moments he dreamed of going on, after fixing Jane Austen, to do the same job on the other major English novelists, then the poets and dramatists, perhaps using computers and teams of trained graduate students, inexorably reducing the area of English literature available for free comment, spreading dismay through the whole industry.	Em momentos faustianos, sonhava em prosseguir, depois de Jane Austen, com o mesmo trabalho sobre os outros grandes romancistas ingleses, e depois os poetas e dramaturgos, utilizando talvez computadores e equipas de estudantes bem treinados, reduzindo inexoravelmente a área da literatura inglesa disponível para a crítica, espalhando o desalento na atividade literária, tomando redutores dezenas de colunas.

Hotovo Internet

- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty
- 4 Příprava textů
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů**
- 7 Různé
- 8 Web jako paralelní korpus
- 9 Přílohy

Extrakce ekvivalentů

– tomu může předcházet:

- zarovnání slov
- označení a zarovnání víceslovných výrazů, větných členů
- syntaktická analýza korpusu (→ treebank)

Překlad s využitím paralelního korpusu

- překladová paměť v systémech podpory překladu
TM – Translation Memory, CAT – Computer-Aided Translation
- překlad podle příkladů
EBMT – Example-Based Machine Translation
- statistický překlad
SMT – Statistical Machine Translation

K tomu všemu se často hodí syntakticky analyzovaný korpus – **treebank**, v našem případě **paralelní treebank**.

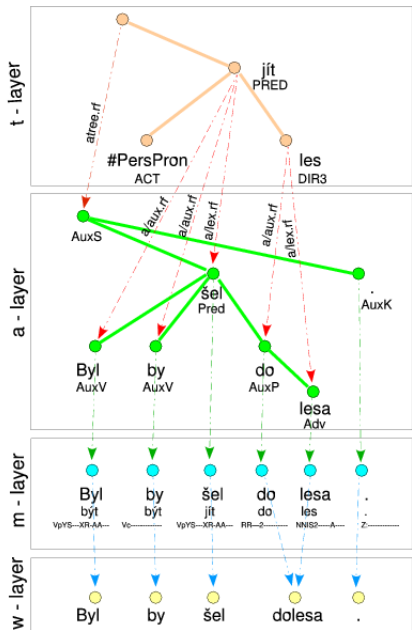
- **Treebanks – databáze stromů**
- Strojový překlad
- Překladové paměti
- Překlad podle příkladů – EBMT
- Statistický překlad
- Příklady

Český treebank

Pražský závislostní korpus 2.0

má více rovin – zhruba podle teorie *funkční generativní popis* (Sgall et al.)

- tektogramatická rovina
- analytická rovina
- morfématická rovina
- rovina grafémů



Paralelní treebanky

- **PCEDT – Prague Czech-English Dependency Treebank**

<http://ufal.mff.cuni.cz/pcedt/>

- ▶ Reader's Digest 1993–1996: 53 000 dvojic vět
- ▶ Wall Street Journal, vybráno z korpusu Penn Treebank: 21 600 dvojic vět

- **PADT – Prague Arabic Dependency Treebank 1.0**

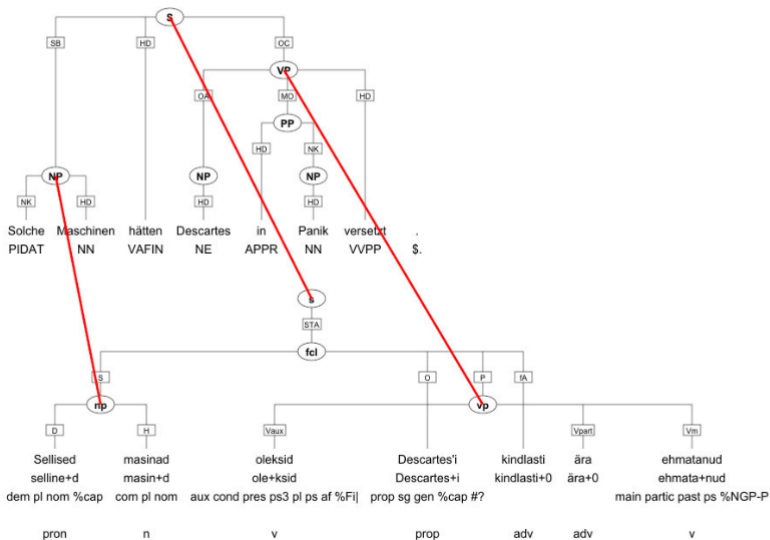
http://ufal.mff.cuni.cz/padt/PADT_1.0/

– zatím jen arabsky

- **Nordic Treebank Network**

<http://w3.msi.vxu.se/~nivre/research/nt.html>

Phrase alignment: example



Ne vždy je možné/nutné analyzovat všecko

– stačí označit některé syntaktické celky, viz korpus OPUS:

```
<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>
```

- Treebanks – databáze stromů
- **Strojový překlad**
- Překladové paměti
- Překlad podle příkladů – EBMT
- Statistický překlad
- Příklady

Google – http://www.google.com/language_tools

The screenshot shows a Mozilla Firefox browser window with the title "Language Tools - Mozilla Firefox". The address bar contains the URL "http://www.google.com/language_tools". The browser's menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Tools", and "Help". The browser toolbar shows navigation icons (back, forward, refresh, stop, home) and a search bar with the same URL. Below the toolbar, there are several bookmarks: "Red Hat, Inc.", "Red Hat Network", "Support", "Shop", "Products", and "Training".

The main content area is titled "Translate" and contains two sections:

- Translate text:** A large empty text input box. Below it is a dropdown menu set to "Spanish to English" and a "Translate" button.
- or**
- Translate a web page:** A text input box containing "http://". Below it is a dropdown menu set to "Spanish to English" and a "Translate" button.

A callout box on the right side of the page contains the text: "Google Toolbar instantly translates words on English web pages into other languages" with a blue link "Download now" below it.

Translate - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://translate.google.com/ti Go

Red Hat, Inc. Red Hat Network Support Shop Products Training

Google Translate

[Get Translation Browser Buttons](#) | [Language Tools](#)

Translate Text

Original text:

Prag ist grundsätzlich an einem langen Tag ausreichend zu besichtigen, man langweilt sich jedoch auch nicht, wenn man eine ganze Woche dort bleibt. Im Sommer ist es so überlaufen, daß es z.T. keinen Spaß mehr macht.

Automatically translated text:

Prague is to be visited in principle on a long day sufficiently, one bores itself however also not, if one remains a whole week there. In the summer it overflowed in such a way that it makes partially no more fun.

German to English Translate

Done

- Treebanks – databáze stromů
- Strojový překlad
- **Překladové paměti**
- Překlad podle příkladů – EBMT
- Statistický překlad
- Příklady

– databáze ekvivalentů, většinou vět a (terminologických) výrazů

Využití:

- opakování vět nebo výrazů uvnitř dokumentu
- opakování vět nebo výrazů v různých dokumentech, různé verze téhož dokumentu
- stejná nebo příbuzná témata, ne nutně technická ("birdwatching")
- originál v elektronické podobě, překlad ve stejném formátu
- čím víc a déle, tím lépe

Výhody:

- využití minulé práce (i cizí)
- dodržení stejné terminologie
- stejné prostředí pro různé formáty

Odkazy:

Déjà Vu: <http://www.atril.com/>

SDL SDLX: <http://www.sdlintl.com/products/sdlx/nav/main.htm>

STAR TRANSIT: <http://www.star-ag.ch/products/>

TRADOS TRANSLATOR'S WORKBENCH: <http://www.trados.com/>

Translation Memory:

http://dmoz.org/Computers/Software/Globalization/Computer_Aided_Translation/Translation_Memory/

More Translation Memory Tools (not many more, but good ones)

by Suzanne Assénat-Falcone

<http://www accurapid.com/journal/12TM.htm>

How To Select the Right CAT Tool Solution

<http://www.languagepartners.com/reference-center/whitepapers/howto.htm>

What you need to know about Translation Memories

<http://www.multilingualwebmaster.com/library/trmemories.html>

- Treebanks – databáze stromů
- Strojový překlad
- Překladové paměti
- **Překlad podle příkladů – EBMT**
- Statistický překlad
- Příklady

Překlad podle příkladů – EBMT

Example-based Machine Translation

- „překlad podle analogie“
- předchozí překlady slouží k překladu nového textu
- jako dvoujazyčný slovník + překlady
- data vydrží déle než teorie

Možnosti:

- holý text
- syntaktická struktura
- kombinace

místo pravidel databáze ekvivalencí mezi výrazy příslušných jazyků – příklady překladů, k tomu je třeba:

- 1 databáze ekvivalencí
- 2 algoritmus, který ke každému výrazu na vstupu vyhledá v databázi nejblížeší výraz
- 3 při hledání se může uplatnit tezaurus s hierarchií, v níž se hledá nejspecifičtější výraz nadřizený oběma porovnávaným
- 4 abstraktní schéma, které bude zaplněno tím, čím se vstup od příkladu v databázi liší

Příklad

Databáze příkladů

wildlife protection – ochrana volně žijících zvířat

radiation protection – ochrana před radiací

police protection – policejní ochrana

Tourists eat hamburgers. – Turisté jedí hamburgery.

Acid eats metal. – Kyselina ničí kov.

Vstup

endangered species protection, tropical forest protection, ozone layer

protection, protection of inhabitants

noise protection, drugs-related hazards protection

government protection, neighbourhood watch protection

She eats a lot of vegetables.

Exhaust fumes eat the marble statues.

- Treebanks – databáze stromů
- Strojový překlad
- Překladové paměti
- Překlad podle příkladů – EBMT
- **Statistický překlad**
- Příklady

Stručný popis statistické metody I

- Překlad z francouzštiny do angličtiny, Brown et al., 1989
- Inspirace z kódování signálu: anglické věty byly zkresleny šumovým kanálem do vět francouzských. Jak najít původní anglické věty?
- Překladem anglické věty S může být kterákoli francouzská věta T . Každé dvojici S a T přisoudíme podmíněnou pravděpodobnost $P(T|S)$, že překladatel přeloží větu S větou T .
- K zadané větě T hledáme nejpravděpodobnější S , která byla přeložena jako věta T .

Stručný popis statistické metody II

- Chceme tedy takovou větu S , která maximalizuje pravděpodobnost $P(S|T)$. Podle Bayesovy věty pak můžeme napsat:

$$P(S|T) = \frac{P(S)P(T|S)}{P(T)}$$

Jmenovatel nezávisí na S , a tak stačí najít takové S , které maximalizuje součin $P(S)P(T|S)$.

- ▶ $P(S)$ pravděpodobnost S v modelu zdrojového jazyka (volba a pořadí slov ve větě S)
- ▶ $P(T|S)$.. pravděpodobnost překladu věty S větou T (jaká slova z S vedla ke slovům v T).

Stručný popis statistické metody III

- Pro systém strojového překladu je tedy třeba:
 - 1 spočítat pravděpodobnosti jazykového modelu
 - 2 spočítat pravděpodobnosti překladového modelu
 - 3 najít takovou větu S, která maximalizuje součin obou pravděpodobností

- Treebanks – databáze stromů
- Strojový překlad
- Překladové paměti
- Překlad podle příkladů – EBMT
- Statistický překlad
- **Příklady**

Strojový překlad literárního textu (systém APAČ) I

CATCH22 26.01.1989 21:12 1

/-1- he found luciana sitting alone at a table in the allied officers' night club, where the drunken anzac major who had brought her there had been stupid enough to desert her for the ribald company of some singing comrades at the bar.-2-

-1- @ našel lucianu, jak sedí osamoceně, na tabulce v nočním klubu spojených důstojníků, kde opilý major anzac, který přiváděl tam ji, byl dosti hloupý, aby opouštěl ji pro oplzlou společnost některých zpívajících soudruhů na tyči.-2-

CATCH22 26.01.1989 21:18 2

/-1- " all right, i'll dance with you, " she said, before Yossarian could even speak.-2-

-1- @ " v pořádku, bude tančit s tebou, " řekla, než yossarian dokonce by mohl mluvit.-2-

CATCH22 26.01.1989 21:23 3

Strojový překlad literárního textu (systém APAČ) II

/-1- " but i won't let you sleep with me. "-2-
 -1- ", ale nenechá tě spát s mnou ".-2-

CATCH22 26.01.1989 21:31 4

/-1- " who asked you ? " Yossarian asked her.-2-
 -1- @-2-
 -2- " kdo se ptal tě ? " yossarian se ptal jí.-3-
 -2- " kdo žádal tě ? " yossarian se ptal jí.-3-
 -2- " kdo se ptal tě ? " yossarian žádal jí.-3-
 -2- " kdo žádal tě ? " yossarian žádal jí.-3-

CATCH22 26.01.1989 21:36 5

/-1- " you don't want to sleep with me ? " she exclaimed with surprise.-2-
 -1- @ " nechce spát s mnou ? " zvolala překvapeně.-2-

CATCH22 26.01.1989 21:41 6

/-1- " i don't want to dance with you ".-2-

Strojový překlad literárního textu (systém APAČ) III

-1- @ " nechce tančit s tebou. "-2-

CATCH22 13.02.1989 11:49 7

/-1- there was only one catch and that was catch - 22, which specified that a concern for one's own safety in the face of dangers that were real and immediate was the process of a rational mind.-2-

-1- byl jen jeden úlovek a to bylo hlava 22, která určovala, že zájem o svou vlastní bezpečnost tváří v tvář nebezpečím, která byla reálná a bezprostřední, byl proces racionální mysli.-2-

-1- byl jen jeden háček a to bylo hlava 22, která určovala, že zájem o svou vlastní bezpečnost tváří v tvář nebezpečím, která byla reálná a bezprostřední, byl proces racionální mysli.-2-

CATCH22 13.02.1989 11:54 8

/-1- orr was crazy and could be grounded.-2-

-1- orr byl bláznivý a by mohl být uzemněný.-2-

CATCH22 13.02.1989 12:03 9

Strojový překlad literárního textu (systém APAČ) IV

/-1- all he had to do was ask ; and as soon as he did, he would no longer be crazy and would have to fly more missions.-2-

-1- @-2-

-2-všechno, co musil dělat, bylo se ptát ;, a jakmile dělal, už by nebyl bláznivý a by musil létat více misí.-3-

-2-všechno, co musil dělat, bylo žádat ;, a jakmile dělal, už by nebyl bláznivý a by musil létat více misí.-3-

CATCH22 13.02.1989 12:10 10

/-1- orr would be crazy to fly more missions and sane if he didn't, but if he was sane he had to fly them.-2-

-1- @ orr by byl bláznivý, aby létal více misí, a rozumné, jestliže nedělal, ale, jestliže byl rozumný musil létat je.-2-

CATCH22 13.02.1989 12:17 11

/-1- if he flew them he was crazy and didn't have to ; but if he didn't want to he was sane and had to.-2-

Strojový překlad literárního textu (systém APAČ) V

-1- @ jestliže létal je byl bláznivý a nemusel ; ale, jestliže nechtěl byl rozumný a musel.-2-

CATCH22 13.02.1989 12:25 12

/-1- yossarian was moved very deeply by the absolute simplicity of this clause of catch - 22 and let out a respectful whistle.-2-

-1- @ yossarian byl pohnut velmi hluboce absolutní jednoduchostí této klauzule hlavy 22 a vydal uctivé zapísknutí.-2-

- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty
- 4 Příprava textů
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů
- 7 Různé**
- 8 Web jako paralelní korpus
- 9 Přílohy

Filmové titulky I

`http://www.opensubtitles.org/`

`http://divxsubtitles.net/`

Filmové titulky II

1 / 00:01:15,708 → 00:01:18,270

My name Borat. I like you.

2 / 00:01:19,037 → 00:01:20,026

I like sex.

3 / 00:01:21,091 → 00:01:22,309

It nice.

4 / 00:01:23,403 → 00:01:25,399

This my country of Kazakhstan.

5 / 00:01:26,205 → 00:01:31,818

It locate between Tajikistan and
Kirghistan,
and assholes, Uzbekistan.

1 / 00:01:14,268 → 00:01:18,949

Moje meno je Borat. Mám vás rád.

2 / 00:01:19,084 → 00:01:19,919

Mám rád sex.

3 / 00:01:21,099 → 00:01:22,299

Je hezký.

4 / 00:01:23,219 → 00:01:25,819

Tohle je moje země, Kazachstán.

5 / 00:01:26,819 → 00:01:31,819

Leží mezi Tádžikistánem,
Kirgistánem
a prdelí světa - Uzbekistánem.

Problémy s formátem vstupu

nat_sample.sxw - OpenOffice.org 1.1.2

Soubor Úpravy Zobrazit Vložit Formát Nástroje Okno nápověda

Předformátovaný text Nimbus Sans L 14

Program·Skype·neobsahuje·žádný·adware,
spyware·ani·malware¶
¶
Žádný·spyware·,adware·ani·malware·Ve
společnosti·Skype·se·pyšníme·tím,·že·nabízíme
produkt,·který·chrání·a·udržuje·vaši·bezpečnost
kdykoli·jste·online,·takže·můžete·být·naprosto·bez
obav·.To·znamená,·že·nebudeme·zobrazovat
nežádoucí·a·vtíravé·reklamy·ani·nedovolíme
žádnému·malwaru·či·spywaru·provozovat·svou
činnost.¶
Co·je·to·adware?¶






Skype·не·содержит·вирусов,·шпионских·и¶
рекламных·программных·модулей¶
¶
Мы·в·Skype·гордимся·тем,·что·наша·продукция¶
стоит·на·страже·информационной¶
безопасности·и·интересов·наших·клиентов.¶
Это·значит,·что·мы·не·размещаем·у·себя¶
ненужную·нашим·пользователям,¶
навязчивую·рекламу·и·следим·за·тем,·чтобы¶
в·наших·продуктах·не·было·вирусов·и¶
шпионских·программных·модулей.¶
Что·такое·рекламные·модули?¶

Strana 1 / 1 Výchozí 100% INSERT STD HYP A1

Místo předmluvy - OpenOffice.org 1.1.2

Soubor Úpravy Zobrazit Vložit Formát Nástroje Okno Nápověda

WW-Plain Text Times New Roma 11

 	<p>"Teď už půjdeš spát?" zeptal jsem se. "Slyšela jsi, co říkal ten pán z Marsu?"</p>	<p>- Ну, теперь ты пойдешь спать? - спросил я. - Ты слышала, что сказал тебе дядя с Марса?</p>	0.0608108
 	<p>"Půjdu. Ale vezmeš mě někdy na Mars?" ~~~~~ "Jestli budeš hodná, poletíme tam v létě."</p>	<p>- Пойду. А ты возьмешь меня на Марс?</p>	-0.127273
	<p>Alenka nakonec usnula a já jsem se opět pustil do práce.</p>	<p>- Если будешь хорошо себя вести, летом туда полетим.</p>	0.084
	<p>Pracoval jsem do jedné hodiny v noci. Najednou tiše zabzučel videofon. Stiskl jsem tlačítko. Hleděl na mě Marfan z vyslanectví.</p>	<p>В конце концов Алиса уснула, и я снова сел за работу. И засиделся до часу ночи. А в час вдруг приглушенно заверещал видеofон. Я нажал кнопку. На меня глядел марсианин из посольства.</p>	-0.15
	<p>"Promiňte, prosím, že vás ruším tak pozdě v noci," omlouval se. "Váš videofon ale nebyl vypnutý, myslel jsem si tedy, že ještě nespíte."</p>	<p>- Извините, пожалуйста, что я побеспокоил вас так поздно, - сказал он, - но ваш видеofон не отключен, и я решил, что вы еще не спите.</p>	0.118605

kundecsar.doc - Microsoft Word

Soubor Úpravy Zobrazit Vložit Formát Nástroje Tabulka Okno Nápověda Acrobat

Prostý text Courier New 12 B II

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

"Prosím tě.	. ضم ملأ هنا "
Pořád nás nutí, abychom vystupovali zadarmo.	تالفح مي قن نأ انم نوع قوتوي . لباقم نودب ، ان اجم قتي سوملا .
Jednou ve prospěch boje proti imperialismu, podruhé k výročí revoluce, potřetí k narozeninám nějakého potentáta, a když nechci, aby nás zlikvidovali, musím se vším souhlasit.	. ةدي دج ةع يرذب نوتاي موي لكو . ةيلاي ريبم الا دض حافكلا لجا نم ةرم ، ةروثلل يونسلا ديعلك رخ او ، كاذ داليمب لفتحن ةيلانلا ةرمل او ظافتح الا تدرأ اذا ، ميظعلا لك رياسأ نأ دبالف ، اعم ةقر فلاب شي .
Nevíš, jak jsem se dnes zas rozčilil.	يلع او طغض اذا مب ةركف يأ كدنع سيل مويلا .
"Copak?"	. "وه ام" ةيضار هتلأس
"Copak?"	تترفح يلحملا سلجملا نم ةأرم ا" ضورفملا نع انرضاحت تادبو فزعلا ةياهنلا يفو ، ضورفملا و فزغن نأ ناجملاب قتي سوم لفح ب انتقلال م ةيبابشلا ةنجلل .
"řekla bez zájmu.	م. كلذو .
"Navštívila nás při zkoušce nějaká referentka z národního výboru a začala nás poučovat, co smíme hrát a co nesmíme, a nakonec nás přinutila, abychom uspořádali zadarmo koncert pro Svaz mládeže, ale to nejhorší je, že zítra musím celý den strávit na jakési pitomé konferenci, kde nás budou poučovat o tom, jak má hudba pomáhat budovat socialismus.	لوط يثقا فوسف - شح ام أوسأ نكي نورثري ، يبغ رم توم يف دغ موي ان يف قتي سوملا رود نع هيف ةيكارتشالا .

Struktura textu stejná jako v originále?

Ne nutně. Jazyky se liší v užívání:

- interpunkce
- dělení na věty
- přímé a nepřímé řeči

Příklad

– Izvinite, požalujsta, čto ja pobespokoil vas tak pozdno, – skazal on, – no vaš videofon ne otključen, i ja rešil, čto vy ešče ne spite.

"Promiňte, prosím, že vás ruším tak pozdě v noci," omlouval se. "Váš videofon ale nebyl vypnutý, myslel jsem si tedy, že ještě nespíte."

Zarovnávání textů s odlišnou strukturou

Předpoklady při zarovnávání:

- 1 shodné nebo nepatrně odlišné pořadí vět v paralelních textech
 - 2 minimum přidaných nebo vypuštěných pasáží
 - 3 většina vět odpovídá 1:1, v jiných případech jsou čísla v $m:n$ nízka
- vše kvůli efektivitě

Příliš často neodpovídá realitě!

Řešení?

- úprava textů před zarovnáním
- načtení textů do databáze, hledání korespondencí bez ohledu na pořadí

Zarovnávání slov, výrazů, větných členů

Předpoklad:

- 1 segmentace/tokenizace v paralelních textech (nezávisle)
- 2 zjišťování korespondencí (zarovnání)

Segmentace ale může záviset na druhém jazyku:

- *patentová přihláška*
- *demande de brevet*
- *Patentanmeldung*
- *domanda di brevetto*

Řešení?

Víceúrovňová segmentace!

Drží se překladatelé co nejvíce originálu?

Záleží na typu textu. V beletrii spíše ne.

Důvody:

- cílový jazyk nemá srovnatelný výraz nebo konstrukci
- překladatel dá ze stylistických důvodů přednost jinému výrazu nebo konstrukci, i když má k dispozici „doslovnější“ variantu
- překladatel se bojí, že udělá chybu, když použije identické výrazové prostředky

A když vypadá překlad podobně jako originál –

– tak může jít o neumělý, nepřirozený, doslovný překlad

Navíc překladatelé někdy chybují

– a některé chyby může odhalit jen velmi dobrý znalec obou jazyků

Co když nemáme paralelní, ale jen srovnatelné texty?

Texty mohou být „neparalelní“ v různé míře:

- stejné věty jsou v textech na jiných pozicích
- texty obsahují jen větší či menší podíl stejných vět
- texty nepojednávají o stejném tématu
- texty nejsou ze stejného oboru

Výsledkem je, že:

- výrazu nelze vždy přiřadit jednoznačný překlad
- ne vždy lze z textů překlad zjistit
- četnosti ekvivalentních výrazů v textech nelze srovnávat

Jak využít srovnatelné texty I

Ale:

- je-li téma stejné, ekvivalentní výrazy se vyskytují ve všech jazycích ve srovnatelném kontextu
- v daném oboru a v určité době se ekvivalentní výrazy vyskytují se srovnatelnou frekvencí

Jsou-li texty ze stejného oboru, na stejné téma a ze stejné doby:

- ekvivalentní výrazy se vyskytují v podobných kontextech
- ekvivalentní výrazy jsou srovnatelně frekventované

Jak využít srovnatelné texty II

Hledání ekvivalentu podle srovnatelného kontextu

- 1 vyhledat slovo S_A s kontextem v jazyce A
- 2 přeložit slova v kontextu S_A pomocí *nějakého* slovníku do jazyka B
- 3 vyhledat kontexty s přeloženými slovy v jazyce B
- 4 hledané slovo S_B je to, které je v těchto kontextech nejčastější

Jinak řečeno... (parafráze) I

K čemu jsou parafráze dobré:

- generování (syntéza) přirozeného jazyka
- sumarizace
- hodnocení systémů strojového překladu
- hodnocení dotazovacích systémů

Využití jednojazykového paralelního korpusu k parafrázování

Emma **burst into tears** and he tried to **comfort** her, **saying things to make her smile**.

Emma **cried**, and he tried to **console** her, **adorning his words with puns**.

Jinak řečeno... (parafráze) II

Postup

- ① zarovnání po frázích (skupinách slov)
- ② *This situation is ... in terms of security*
- ③ *under control* → *unter Kontrolle*
- ④ *unter Kontrolle* →
 - ▶ *in check*
 - ▶ *checked*
 - ▶ *curbed*
 - ▶ **curb*
 - ▶ **limit*
 - ▶ **slow down*

(Bannard & Callison-Burch, ACL 2005)

- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty
- 4 Příprava textů
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů
- 7 Různé
- 8 Web jako paralelní korpus**
- 9 Přílohy

Zdroje paralelních textů na webu

Hotové paralelní korpusy

- s webovým vyhledávacím rozhraním (Kačenka, SNK, COMPARA, OPUS)
- přístupné k dalšímu využití (Multext, Acquis Communautaire)

Elektronicky čitelné texty ve více jazycích

- beletrie
(<http://www.logoslibrary.eu>, ...)
- zákony

Web jako korpus?

McEnergy & Wilson (1996):

Korpus je sbírka textů, která

- obsahuje vzorky rozsáhlejších textů
- je reprezentativní
- je konečně velká
- je strojově čitelná
- lze na ni odkázat standardním způsobem

Ale:

- korpus díla Karla Čapka
- trénovací korpusy pro systémy zpracování přirozeného jazyka

neobsahují vzorky, nejsou reprezentativní, nelze na ně odkázat

Proč tedy web nemůže být taky korpus?

Hledání textů na webu ve více jazycích

- 2,6 mld IP adres, z toho 5,1 mil. českých
- 2003: 520 mil. slov česky, 7 mld slov německy, 77 mld slov anglicky (Alta Vista, dolní odhad)

Ručně nebo automaticky?

- automatické metody nutné k získání většího než minimálního množství textů
- úspěšnost může být např. 99 % v přesnosti a 97 % v pokrytí [Ma & Liberman(1999)]
- nezávislé na konkrétních jazycích, výjimky:
 - ▶ substituční pravidla k hledání adres odpovídajících stránek
 - ▶ překladové slovníky k porovnání obsahu stránek
 - ▶ data k identifikaci jazyka (slovník nebo max. 100 000 znaků textu k natrénování identifikátoru)

Postup

- 1 hledání stránek (dokumentů), které mohou být také v jiném jazyce
- 2 hledání překladových ekvivalentů stránek
- 3 filtr: odstranění chybných ekvivalentů

Krok 1: hledání stránek ve více jazycích

- přes odkazy na stránky v různých jazycích na nadřazené stránce
- přes odkaz na překlad stránky
- stránky v určité doméně

Krok 2: hledání překladového ekvivalentu stránky

- s odkazy na překlady snadné
- porovnávání adres stránek (URL) (<http://cs.wikipedia.org/> vs. <http://de.wikipedia.org/>):
 - ▶ ručně vytvořená substituční pravidla (en → cs / big5 / ...)
 - ▶ řetězce označující jazyk často začínají nebo končí charakteristickými znaky: _, -, mohou se v adrese objevit i 2x
 - ▶ Levenštejnova editační vzdálenost (*edit distance*)
 - ▶ ale pozor: <http://de.wikipedia.org/wiki/Zajíc> neodpovídá <http://de.wikipedia.org/wiki/Zajíc>
- porovnávání délky dokumentů, předpoklad: konstantní poměr znaků mezi určitými dvěma jazyky
- na základě automatického zjištění jazyka dokumentu
 - ▶ automatická identifikace jazyka dokumentu
 - ▶ vytvoření všech možných dvojic dokumentů
 - ▶ odstranění nevyhovujících dvojic dokumentů (filtr)

der Feldhase a Jan Zajíc



Krok 3: filtrování

- strukturní filtr: porovnávání HTML značek, případně doplněných údajem o délce příslušného úseku textu
- jazykový filtr: automatická identifikace jazyka
- obsahový filtr: překladový slovník, *cognates*, *anchors*; sekvenční porovnání nebo porovnání automaticky vygenerovaných indexů
- délkový filtr I: znaky (konstantní poměr), odstavce (identita)
- délkový filtr II: likvidace velmi krátkých textů (kratší než 500 znaků)
 - snižují kvalitu korpusu

Problémy I

Málo jazyků, málo dat

- automaticky se z webu získaly paralelní korpusy zatím jen pro málo jazyků (angličtina – francouzština, čínština, arabština, ...)
- obrovský nepoměr mezi angličtinou a ostatními jazyky
- situace se zlepšuje (1997: jen 1 promile adres obsahuje stránky ve více jazycích, ale např. v doméně .de je 10 % německo-anglických adres)

Problémy II

Autorské právo

- šíření textů třetích osob teoreticky vyžaduje jejich souhlas
- lze obejít vystavením adres dokumentů místo dokumentů samotných
- ale pak nelze vystavit zarovnané texty
- adresy i jejich obsah se mění – lze vyřešit využitím internetových archivů

Nevyváženost

Problémy III

Strukturní filtr někdy nepomáhá

- překlady mohou mít jinou strukturu
- v mnoha dokumentech chybí strukturní značkování

Řešení: obsahový filtr (překladač slovník), délkový filtr

Prolézání celé sítě je náročné

Řešení: internetové archivy, např. <http://www.archive.org> (2003: 120 TB, 10 mld stránek)

Stačí-li nám jen něco:

Některé servery vydávají např. zprávy ve více jazycích. Stálý přísun!

Odkazy

- BITS [Ma & Liberman(1999)]
- PTMiner [Chen & Nie(2000)]
- STRAND <http://umiacs.umd.edu/~resnik/strand>
[Resnik & Smith(2003)]



Chen, J. & Nie, J.-Y. (2000).

Automatic construction of parallel English-Chinese corpus for cross-language information retrieval.

In Proceedings of the Sixth Conference on Applied Natural Language Processing, pages 21–28, Seattle.



Ma, X. & Liberman, M. (1999).

BITS: a method for bilingual text search over the web.

In Proceedings of Machine Translation Summit VII. National University of Singapore.



Resnik, P. & Smith, N. A. (2003).

The Web as a parallel corpus.

Computational Linguistics, **29**(3), 349–380.



- 1 Úvod
- 2 Existující korpusy a zdroje dat
- 3 Technické aspekty
- 4 Příprava textů
- 5 Hledání v paralelních korpusech
- 6 Další využití paralelních korpusů
- 7 Různé
- 8 Web jako paralelní korpus
- 9 Přílohy**

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)

<p> "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.</p>

<p> "Všššc!" sykl pan Dursley nahlas.</p>

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

▶ postup

po očíslování odstavců (.txt1)

`<p id="22">"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.</p>`

`<p id="23">"Všššc!" sykl pan Dursley nahlas.</p>`

`<p id="24"> Kočka se ani nepohnula, jenom se na něj přísně podívala. Pan Dursley chvílku uvažoval, jestli se kočky takhle chovají normálně. Zatímco se nutil ke klidu, otevřel si domovní dveře; ještě pořád nehodlal manželce nic říkat.</p>`

▶ postup

po označení českých vět (.txt1)

```
<p id="22">"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.</p>
```

```
<p id="23"><s id="23.1">"Všššc!" sykl pan Dursley nahlas.</s></p>
```

```
<p id="24"> <s id="24.1">Kočka se ani nepohnula, jenom se na něj přísně podívala.</s> <s id="24.2">Pan Dursley chvílku 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></p>
```

po zarovnání (.seg)

```
<p id="22"><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></p>
```

```
<p id="23"><s id="23.1"><seg id="89">"Všššc!" sykl pan Dursley nahlas.</seg></s></p>
```

```
<p id="24"> <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 chvílku 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></p>
```

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

