

# Gradnja in raba večjezičnih korpusov

Univerza v Ljubljani  
Filozofska fakulteta  
Oddelek za prevajalstvo

Leksika slovenskega jezika

Študijsko leto 2009/10



asist. dr. Darja Fišer

## Pregled predelane snovi

### Teme:

### Spretnosti:

- ✓ raba besed & vloga sobesedila
- ✓ opazovanje & določanje pomena
- ✓ besede v besednih zvezah
- ✓ besede v besedilih
- ✓ besede in kultura

- ✓ konkordance
- ✓ besedne skice
- ✓ podkorpusi
- ✓ beseni seznami
- ✓ gradnja korpusov

# Pregled današnjega predavanja

## Teme:

- ✓ vzporedni korpusi
- ✓ primerljivi korpusi
- ✓ načela gradnje
- ✓ postopek izdelave
- ✓ uporaba večjezičnih korpusov

## Spretnosti:

- ✓ predobdelava
- ✓ stavčna poravnava
- ✓ označevanje
- ✓ iskanje
- ✓ terminografija

3

## Vzporedni korpusi

Baker (1995)

- ▶ izvirnik & prevod(i)
- ▶ stavčno poravnani
- ▶ včasih zajeti zgolj prevodi brez izvirnika  
(npr. tehnična dokumentacija)
- ▶ včasih o izvirniku sploh ne moremo govoriti  
(npr. EU)

4

# Primer vzporednega korpusa

## IJS-ELAN Slovene part

Search for "telo" as PARA

69 hits

---

[parl](#): Predsednik in podpredsedniki so za svoje delo odgovorni državnemu zboru. Kolegij predsednika državnega zbora Kolegij predsednika je posvetovalno **telo** predsednika državnega zbora. Kljub temu, da je funkcija tega telesa zgolj posvetovalne narave, pa ima kolegij

The collegium of the President is the advisory body to the President of the National Assembly.

---

[vade](#): belih krvničk ( granulocitov, limfocitov T ) in aktivirajo njihovo fagocitozno sposobnost ter tako ovirajo vdor mikroorganizmov v **telo** in pospešujejo njihovo uničevanje. Dokazano je tudi protivirusno delovanje ameriškega slamnika na viruse gripe in herpesa. -

By increasing the number of leukocytes ( granulocytes, lymphocytes T ) and activating their phagocytosis ability, they inhibit the penetration of microorganisms in the body, and accelerate their extirpation.

---

[orwl](#): je, koliko mehkejše je njeno **telo** zdaj, ko ni imela več pasu.

He noticed how much softer her waist seemed to feel now that the sash was gone.

---

# Vzporedni korpusi za slovenščino

## ► Evrokorpus

- korpus prevodov zakonodaje EU
- ang-slo, nem-slo, fra-slo, ita-slo, špa-slo
- neoznačen

## ► nl2.ijs.si

- SVEZ-IJS
  - označen ang-slo del Evrokorpusa, 10 mio besed
- ELAN
  - 15 besedil z različnih področij, 1 mio besed
- TRANS
  - za terminografsko delo, 900.000 besed

# Primerljivi korpusi

Baker (1995)

## Primerljivi:

- ▶ ne vsebujejo izvirnikov & prevodov, temveč besedila v različnih jezikih, ki so si med seboj podobna glede na žanr, področje ipd.
- ▶ niso stavčno poravnani

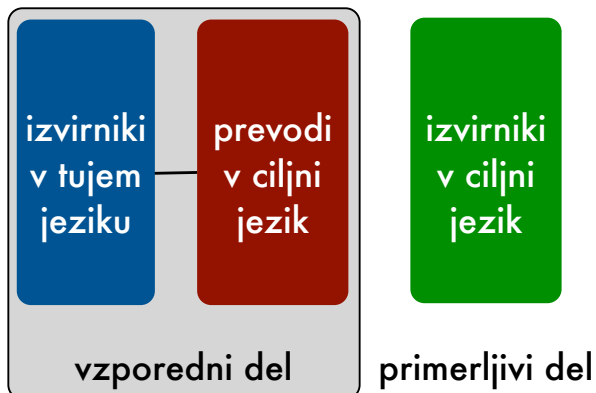
## Prevodoslovni:

- ▶ vsebujejo izvirnike in prevedena besedila v istem jeziku
- ▶ enojezični primerljivi korpusi

7

## Razvoj primerljivih korpusov za slovenščino

- ▶ v okviru projekta Slovensko prevodoslovje (Vintar)



8

# Raba večjezičnih korpusov

## ► pomoč pri učenju tujih jezikov:

- opazovanje rabe besed v tujem jeziku
- iskanje besedišča, ki ni zajeto v slovarjih  
(npr. novo besedišče, terminologija)

## ► vir pri prevajanju:

- iskanje prevodnih ustreznic

## ► zbirka podatkov za prevodoslovne, kontrastivne & jezikovnotehnološke raziskave:

- proučevanje prevajalskega procesa, prevajalskih norm, vpliva družbenega konteksta na prevode, jezika prevodov, razvoj jezikovnih aplikacij, strojno prevajanje

9

# Raba večjezičnih korpusov

Display: ☒ Bilingual ☐ KWIC ☐ Word List

Context: ☐ 10 ☐ 20 ☒ 40 ☐ 80 ☐ 160

Corpus: ☒ SVEZ-IUS-SL ☐ SVEZ-IUS-EN

☐ ELAN-SL ☐ ELAN-EN

☐ TRANS2-SL ☐ TRANS2-EN

Corpus Query:

On aligned:

☐ require ☒ forbid

Seznam, ki ga podpiše **predsednik**, se pred glasovanjem

The list, to be signed by the chairman, shall be made available to the participants of the general meeting for inspection before voting

Seznam, ki ga podpiše **predsednik**, se pred glasovanjem

The list, to be signed by the chairman, shall be made available to the participants of the general meeting for inspection before voting

je dr. Janez Drnovšek, **predsednik** stranke, štirikrat

Janez Drnovšek, the leader of the LDS, participated in an on-line chat four times during the 2000 election campaign

Po drugi strani je **predsednik** spregledal dejstvo, da

On the other hand, Drnovšek overlooked the fact that the on-line chat includes several participants, which reduces the transfer of authority

vprišanje o tem je **predsednik** vlade odgovoril: " Da

When questioned about this, the Prime Minister answered

da sta Slovenija in **predsednik** vlade pripravljena

Thus, the site serves to not only introduce the PM but also argue that Slovenia and the PM are quite ready to accept the challenges of Europe

10

# Načela gradnje

Biber (1993)

1. tip korpusa  
(namen)
2. velikost/reprezentativnost  
(vzorčenje)
3. avtorske pravice  
(dovoljenje)

11

# Postopek izdelave

- 1.pridobivanje besedil  
(klasični viri, internet, pomnilniki prevodov)
- 2.predobdelava besedil  
(OCR, poenotenje zapisa, čiščenje besedil)
- 3.označevanje korpusa  
(večjezično, lematizacija, oblikoskladenjsko)
- 4.stavčna poravnava  
(ročna validacija: WinAlign, avtomatska)
- 5.uporaba konkordančnika

12

# Stavčna poravnava



izvirnik

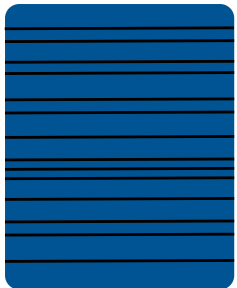


prevod

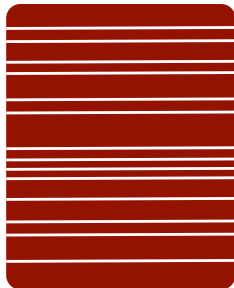
13

# Stavčna poravnava

1. segmentacija



izvirnik

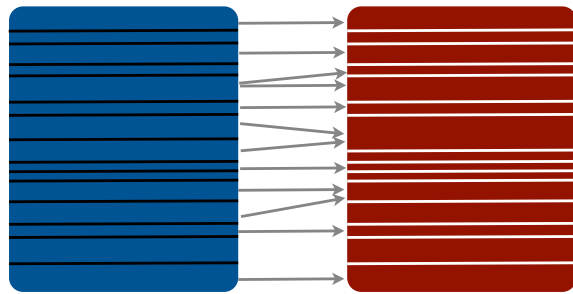


prevod

14

# Stavčna poravnava

## 2. poravnava



izvirnik

prevod

15

## Orodja za stavčno poravnavo

### ► WinAlign

- Trados
- Windows
- plačljiv
- ročno popravljanje

### ► DVX

- Atril
- Windows
- plačljiv
- ročno popravljanje

### ► Hunalign

- Linux & Windows
- brezplačen
- avtomatski

### ► Uplug

- Linux & Windows
- brezplačen
- avtomatski & ročno popravljanje

16



# Orodja za iskanje

## ► ParaConc

- plačljiv, na voljo demo različica
- stavčna poravnava
- iskanje z regularnimi izrazi
- urejanje konkordanc
- iskanje kolokacij
- ne podpira znakovnega nabora Unicode

17

# Zapomnite si

- vzporedni korpusi vsebujejo izvirnike & prevode v 1 ali več jezikov
- primerljivi korpusi vsebujejo besedila v 1 ali več jezikih podobnih žanrov, področij ipd.
- vzporedne korpuse moramo pred uporabo stavčno poravnati
- vzporedne korpuse uporabljamo za:
  - opazovanje rabe besed
  - iskanje prevodnih ustreznic
  - iskanje terminologije
  - znanstvene raziskave

18

# Vaje z vzporednimi korpusi

1. S pomočjo Evrokorpusa raziščite prevodne ustreznice za angleški izraz »practices«. Ali se izraz v različnih besednih zvezah prevaja različno?
2. V korpusu ELAN raziščite, ali se izraz »korak« prevaja še kako drugače kot »step«. V katerih primerih pride do spremembe besedne vrste?
3. Na svetovnem spletu ali iz lastnega arhiva poiščite poljubno izvorno besedilo in njegov prevod ter ju avtomatsko poravnajte na ravni stavkov z orodjem ParaConc.
4. Izberite si 5 poljubnih besed in besednih zvez v izvornem jeziku in s pomočjo vzporednih konkordanc najдите prevodne ustreznice zanje. Svoje ugotovitve primerjajte z informacijami v splošnem dvojezičnem slovarju.

19

## Za radovedne

- ▶ Baker, Mona, 1993: Corpus Linguistics and translation studies: Implications and applications. V: Text and Technology. 17–45.
- ▶ Biber, Douglas, 1993: Representativeness in Corpus Design. Literary and Linguistic Computing 8/4. 243–257.
- ▶ Kenny, Dorothy, 2001: Lexis and Creativity in Translation: A corpus-based study. Manchester: St Jerome.
- ▶ Olohan, Maeve, 2004: Introducing Corpora in Translation Studies. London: Routledge.
- ▶ Tymoczko, Maria, 2000: Translation and political engagement: Activism, social change and the role of translation in geopolitical shifts. The Translator 6. 23–27.
- ▶ Vintar, Špela, 2008: Corpora in Translation: A Slovene Perspective. Journal of Specialized Translation. Issue 10.

20