

Paralelaj korpusoj

Sonja Petrović Lundberg

Enhavo

- Pri korpusoj ĝenerale
- Kial krei paralelkorpusojn?
- Kiel krei paralelkorpusojn?
- En-Eo-korpuso
- Diskuto

Pri korpusoj

Kolekto de lingvodatumoj:

- unueca / samtipa
- strukturita
- serĉebla

Pri korpusoj

Enhavspeco:

- Tekstaroj (Libroj, retmesaĝoj, protokoloj...)
- Voĉregistraĵoj
- Kombinoj/aliaj

Pri korpusoj

Enhavformato:

- Simpla teksto
- XML
- aliaj

Pri korpusoj

Aldonaj informoj:

- Fonto, aŭtoro, dato...
- Gramatika analizo
- Traduko

Pri korpusoj

Lingvoj:

- unu (ev. korekta k. nekorekta, aŭ voĉo k. teksto k.a.)
- pluraj

Unulingvaj korpusoj

- Ekzemploj de aŭtenta lingvouzo
- Statistiko pri vortofteco, vortkunapero k.s.
- Utilaj por lingvolernado, verkado de vortaroj, terminaroj, gramatiklibroj k.t.p.

Unulingvaj korpusoj

- <http://tekstaro.com/> - 4,7 milionoj, multaj verkoj, XML
- <http://corp.hum.sdu.dk/cqp.eo.html> - 58,4 milionoj, reto kaj Vikipedio, gramatike analizitaj tekstoj
- EPAK - 100 000 vortoj

Paralelaj korpusoj

- Sama enhavo en pluraj lingvoj.
- Ligoj inter la diverslingvaj versioj:
 - dokumentnivele
 - paragrafnivele
 - fraznivele
 - vortnivele

Kial krei paralelkorpusojn?

- Lingvokomparado
- Maŝintradukado
- Aŭtomata vortarkreado
- Helpo en kreado de ontologioj,
terminaroj, materialo por
lingvolernantoj...

Uzeksemplo 1

1. Kiel esprimi nacilingvan nociojn en E-o?
2. Paralelkorpuso montras kiel la aliaj faris.
3. Statistikoj montras la plej uzatan alternativon.
4. Konteksto montras la plej taŭgan alternativon.

Uzekzemplo 2

1. Bezonatas terminologia vortaro pri certa temo.
2. Aŭtomata analizo de tekstaro pri tiu temo donas terminkandidatojn.
3. Analizo de paralela korpuso montras terminuzon en diversaj lingvoj kaj tradukkkandidatojn.

Kiel krei paralelkorpusojn?

1. Decidu celon.
2. Difinu tekstogenron kaj formaton.
3. Kolektu tekstojn.
4. "Purigu" kaj unuecigu.
5. Laŭbezone aldonu informojn.
6. Segmentigu kaj paraleligu.

Paralelig-metodoj

- ◉ Fraznivele:

- ◉ Longecbazita – nombro de literoj
aŭ vortoj

- ◉ Vortarbazita

- ◉ Kombinita (Hunalign) – preta
vortaro aŭ provizora aŭtomata

- ◉ GIZA++ por vortnivela ligado

Ekzistantaj paralelkorpusoj

- EuroParl, Kanada parlamento, UN
- Uzinstrukcioj
- Filmsubtitoloj
- OPUS – <http://urd.let.rug.nl/tiedeman/OPUS/>
- Uplug – <http://urd.let.rug.nl/>

Anglo-Esperanta parallelkorpuso

- 2,3 milionoj da vortoj
- Morfosintakse analizitaj frazoj
- Fraznivelaĵoj
- Literaturo (ĉ. 73%), Martinus (13%), UNESCO (4%), politikaj dok (3%), cetero (7%)
- Projekto Gutenberg, eLIBREJO, reto, UEA, MI

Terminologia studeto

- 300 000 vortoj pri Martinus-kosmologio
- 2 000 000 vortoj alitemaj
- 300 terminkandidatoj elfiltritaj laŭ diversaj metodoj
- Ĉ. 50% validaj terminoj
- Sekva paŝo - tradukoj

Diskuto

- Ĉu indas?
- Eblaj aplikoj?
- Alternativaj korpusoj?
- Rutinoj por kreskigi korpuso(j)on?

Kie elŝuti?

- <http://code.google.com/p/paralela-korpuso/>
- Se vi bezonas aliajn dosierojn aŭ helpon, ne hezitu kontakti min, sonja ĉe ikso.net.