Pregled korpusov za slovenščino

Nataša Logar

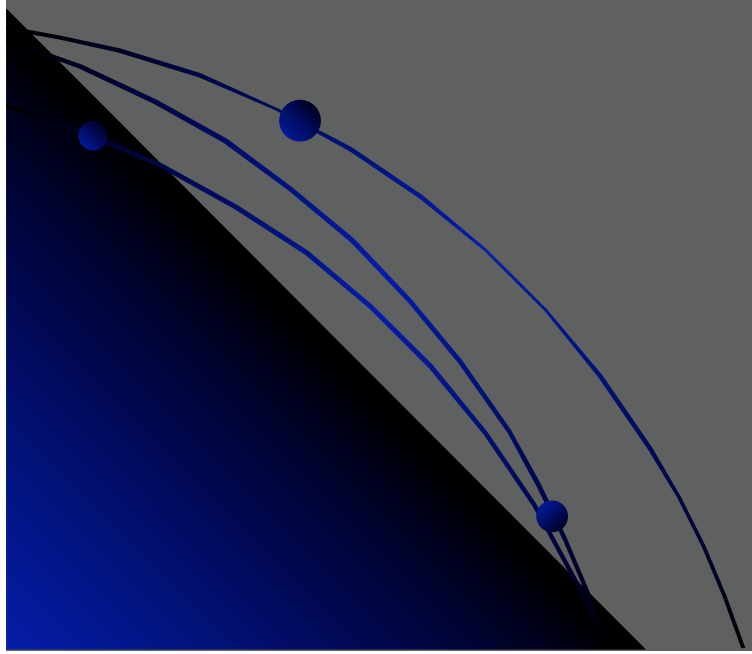
3. posvet Slovenskega društva za jezikovne tehnologije
17. 10. 2008

1 Korpusi po tipu

* Pregled je okvirni.

** Navedena je le vodilna ustanova in/ali posameznik.

*** Notranja razvrstitev je naključna.



Korpusi splošnega jezika

- Korpus slovenskega jezika FIDA

  www.fida.net, 100 mio, 1994-2000, reprezentativni

  FF UL, DZS, IJS, Amebis

- Korpus slovenskega jezika FidaPLUS

  www.fidaplus.net, 621 mio, 1990-2006, reprezentativni

  Jezikovni viri za slovenščino (Marko Stabej)

  FF UL

  > Oblikoskladenjsko označevanje: korpus jos100k, jos1M
  http://nl.ijs.si/jos/ (Tomaž Erjavec)

- Beseda

  http://bos.zrc-sazu.si/main_si_l2.html, Inštitut za slovenski jezik Frana Ramovša, 3 mio, 1858-1996 (112 del leposlovja), Primož Jakopin

- Nova Beseda

  http://bos.zrc-sazu.si/s_beseda.html

  Inštitut za slovenski jezik Frana Ramovša, 240 mio (od tega 169 mio Delo), besedila iz pretežno zadnjih 10 let, Primož Jakopin

Govorni korpusi

- Učni korpus govorjene slovenščine

  http://torvald.aksis.uib.no/talem/jana/s9.html (geslo), Univerza v Bergnu, Norveška, 89 min (15.000 besed), Jana Zemljarič Miklavčič (FF UL)

Govorni korpusi, namenjeni predvsem razvoju govornih aplikacij, sistemov za samodejno razpoznavanje spontanega govora:

- SiBN

  podnapisi dnevnoinformativnih oddaj RTV Slo (teletekst), 2003-04, 1.300 oddaj, 2,3 mio, FF UL + FE UL, Grega Milharčič

- SpeechDat II

  1997, 1000 govorcev > baza 43 izgovorjav vsakega govorca v dolžini okr. 5 min, posneto na telefonskem omrežju (lastnik baze: Siemens AG), FERI UM, Janez Kaiser

- **GOPOLIS**

  povpraševanje po letalskih informacijah, 15 ur (projekt: SQEL), Jerneja Gros (Alpineon)

- **Turdis, Turdis-1, Turdis-2**

  poizvedbe po turist. inf. v agencijah in hotelih, cilj: 200 dialogov

  (T-1: 106 min, 15.000; T-2: 214 min, 32.000), FERI UM, Darinka Verdonik

- **BNSI Broadcast News**

  dnevnoinf. oddaje RTV Slo (osrednji del: 1999-2003, 36 ur), FERI UM, Andrej Žgank

- SloParl

  razprave v Državnem zboru RS (govorni del: 100 ur, 255 govorcev, 650.000 besed; pisni del (prepisi): 23 mio besed), FERI, Andrej Žgank

> Več o govornih tehnologijah na FERI gl. 1. posvet SDJT, Zdravko Kačič, http://www.sdjt.si/dogodki/MB2007/KACIC-jezikovni-viri-2007.pdf

Vzporedni korpusi

- Korpus TELRI ‘Plato’

  1998, http://nl.ijs.si/telri/Republic/, 10 jezikov, Platonova Republika, Inštitut za slovenski jezik, Primož Jakopin

- MULTITEXT-East (G. Orwell: “1984”)

  1998, http://nl.ijs.si/ME/, več jezikov, 100.000 besed, IJS, Tomaž Erjavec, zadnja različica 3 (2004)

  > 1/3 (30.000 besed): površinskoskladenjsko označeni korpus SDT (Slovenska odvisnostna drevesnica), Nina Ledinek (Inštitut za slovenski jezik Frana Ramovša)

- IJS-ELAN

  http://nl.ijs.si/elan/, http://nl2.ijs.si/index-bi.html, SLO-AN, 1 mio, 15 besedil, IJS, Tomaž Erjavec

- TRANS

  http://nl2.ijs.si/index-bi.html, AN-SLO, 2002/03-03/04, 100 besedil različnih strokovnih področij (2 mio), FF UL, Špela Vintar

- SVEZ-IJS

  http://nl.ijs.si/svez/, http://nl2.ijs.si/index-bi.html, AN-SLO, 10 mio, prevodi zakonodaje EU (ACQUIS), Jasna Belc (SVEZ), Tomaž Erjavec (IJS)

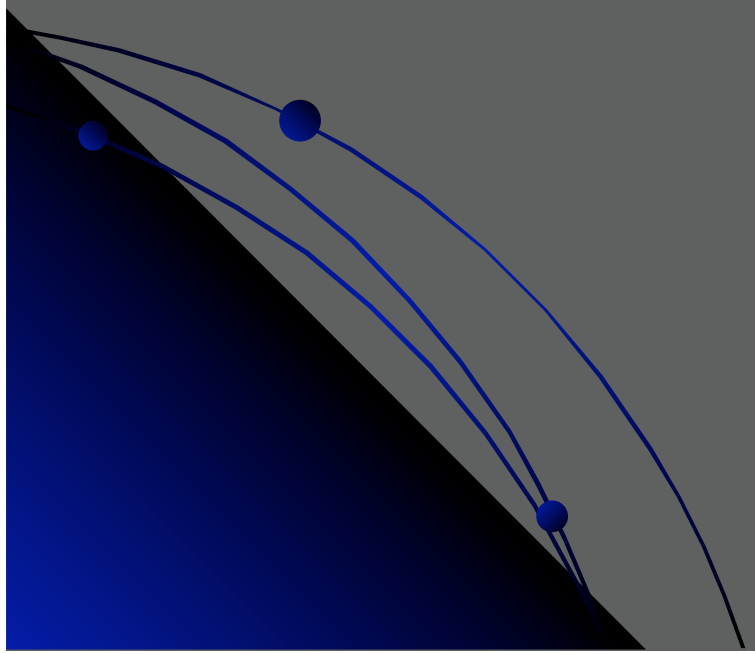
- Evrokorpus

  http://evrokorpus.gov.si/, razširjeni korpus: 22 > 23 jezikov (211 mio), prevodi pravnih aktov EU, prevodi slovenske zakonodaje, SVEZ, Miran Željko

Specializirani korpusi

- Verbalni napadi na JNA (VAYNA)

  http://nl.ijs.si/et/talks/korpus/vayna-hdr.html, 360 časop. člankov april-avgust 1988, 260.000, IJS, Peter Tancig

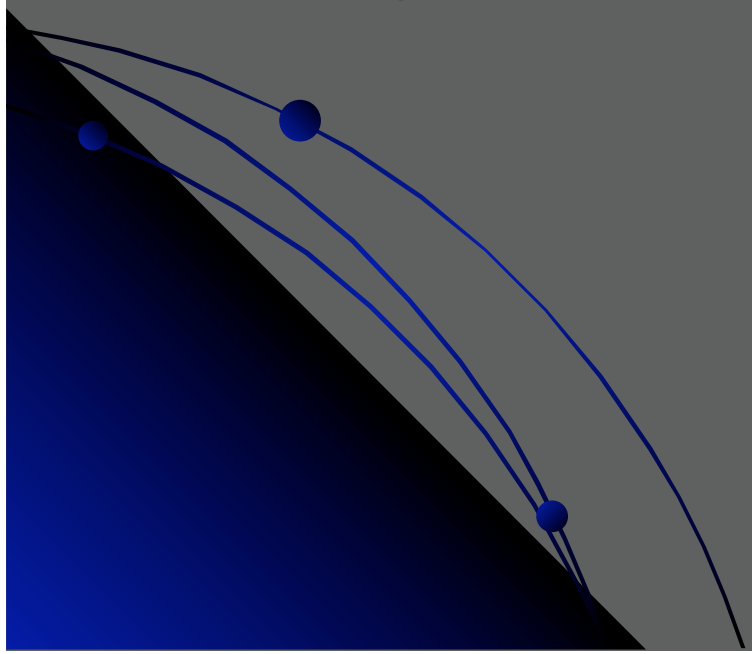


- Korpus Dnevi slovenske informatike

  http://nl2.ijs.si/dsi.html, 2003-08, 1,4 mio, Islovar (http://www.islovar.org/iskanje_enostavno.asp)

  > iFpX/iKorpus

  korpus DSI + računalniški del korpusa FidaPLUS, 14 mio, Špela Vintar



- Korpus vojaških besedil Grizold

  http://ksvjg.fdv.uni-lj.si (geslo), 5,5 mio, FDV UL, Nataša Logar

- Korpus besedil odnosov z javnostmi

  http://www.korp.fdv.uni-lj.si/, 1,8 mio, FDV UL, Nataša Logar

- Večjezični korpus turističnih besedil

  SLO-AN-IT, 30 mio, ZRS UP, Vesna Mikolič

2 Nekaj ustanov

(po abecednem redu)

- Amebis, d. o. o., Kamnik
- Fakulteta za družbene vede Univerze v Ljubljani (zlasti Center za družboslovnoterminološko in publicistično raziskovanje)
- Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru (zlasti Laboratorij za digitalno procesiranje signalov)
- Fakulteta za humanistične študije Univerze na Primorskem (zlasti Oddelek za slovenistiko)
- Filozofska fakulteta Univerze v Ljubljani (zlasti Oddelek za slovenistiko, Oddelek za prevajalstvo)
- Inštitut "Jožef Stefan" (zlasti Odsek za tehnologije znanja, Odsek za inteligentne sisteme)
- ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša (zlasti Laboratorij za korpus slovenskega jezika)

+ posamezniki

3 Že zastavljena prihodnost

- francosko-slovenski vzporedni korpus (FF UL: M. Schlamberger Brezar)
- italijansko-slovenski vzporedni korpus ISPAC (FF UL: T. Mikolič Južnič)
- japonsko-slovenski vzporedni korpus (FF UL: K. Hmeljak Sangawa)
- korpus usvajanja slovenščine kot drugega/tujega jezika (FF UL: Mojca Stritar)
- korpus usvajanja angleškega jezika (FHŠ UP: Neva Čebron)
- "Sporazumevanje v slovenskem jeziku", referenčni korpus z govornim podkorpusom, http://www.slovenscina.eu/Vsebine/Domov/Domov.aspx (Amebis: Miro Romih, S. Krek)
- ...

Popravki in dopolnitve pregleda:

natasa.logar@fdv.uni-lj.si