Orodja za gradnjo korpusov

Posvet SDJT

17.10.2008

Kako do korpusa - DIY

- “resni” vs. “garažni” korpusi
- “Kdor ne korpusa, ni Slovenc.”
  - korpusati – izdelovati, obdelovati ali uporabljati korpuse pri (jezikoslovnih) raziskavah in konjičkih
- orodja za gradnjo vs. orodja za pregledovanje

Orodja za gradnjo

- ...namen? ...obseg? ...tip? ...drugi parametri?
- 1. scenarij: besedila z interneta
  - WebBootCat
- 2. scenarij: besedila z drugih virov

WebBootCat

- del korpusne orodjarne SketchEngine (http://www.sketchengine.co.uk/)
- avtomatizirana gradnja specializiranega korpusa
  - osnova za gradnjo: ključne besede
  - s pomočjo spletnega iskalnika
  - samodejno čiščenje besedil
  - tokeniziranje
  - oblikoskladenjsko označevanje
    - Le za nekatere jezike, ne tudi slovenščino
- kot BootCat (Perl) in JBootCat (Java) na voljo tudi za lokalno namestitev

user: Špela Vintar, **free space:** 436774 tokens

This application uses the Yahoo! Web Services .

Please make sure you have JavaScript enabled in your web browser.

Seed words ribištvo ribarjenje "ribolovne cone" staleži

Use space as separator. Enclose multiwords expressions into quotes (").

Brskaj ...

Upload seed words in a plain text file -- one expression per line.

Language Slovenian

Select the language of the corpus to be built.

Select URLs ☐

Check this box if you would like to manualy filter found URLs before the webpages will be downloaded.

Tag corpus ☐

Your corpus will be POS-tagged and lemmatised using the TreeTagger . Following languages are currently supported: Bulgarian, Dutch, English, French, German, Italian, Russian, Spanish. This option has no effect if used with any other languages.

user: Špela Vintar, **free space:** 436774 tokens

My corpora

Build new corpus

Corpus name	Language	Tokens	POS-tagged	
ribistvo	Slovenian	63232	no	[open] [downl raw] [downl vert] [delete]

Home | Concordance | Word List

View options | Sample | Filter | Sort | Frequency | Collocation | Save

Page 1 of 3 Go Next | Last

00002 vlečna mreža - pridnena vlečna mreža za **ribolov** senegalskega osliča - Uporaba dvojne vreče
00002 2006 | Število plovil z dovoljenjem za **ribolov** | 36 | 36 | 36 | 36 | 36 | Predujem v EUR
00002 2006 | Število plovil z dovoljenjem za **ribolov** | 31 | 31 | 31 | 31 | 31 | Predujem v EUR
00002 1 Ribolovna cona , v katerem je dovoljen **ribolov** z živimi vabami : Severno od zemljepisne
00002 Minimalna dovoljena velikost očesa mreže za **ribolov** z živimi vabami : 8 mm . 7 . 3 V skladu
00002 Število plovil z dovoljenjem za sočasni **ribolov** | 15 | 15 | 15 | 15 | 15 | Pristojbine
00002 VELJAVNOST 1 . Izpolnjevanje pogojev za **ribolov** 1 . 1 Vsako plovilo , ki želi ribariti
00002 sporazumom , mora izpolnjevati pogoje za **ribolov** v ribolovni coni Mavretanije . 1 . 2 Da
00002 podpisati kapitan plovila . Za plovila za **ribolov** izrazito selivskih vrst se uporablja poglavje
00002 tehničnih pregledov , ki veljajo za plovila za **ribolov** tuna , plovila s površinskimi parangali
00002 II Priloge I se dovoljenja za plovila za **ribolov** tun izdajo za obdobje 12 mesecev . Izvirnik
00002 , vnese v seznam plovil z dovoljenjem za **ribolov** , ki se pošlje mavretanskim organom ,
00002 na poglavje VIII te priloge si plovila za **ribolov** tuna s plavarico prizadevajo , da zaposlijo
00002 na plovilo , medtem ko morajo plovila za **ribolov** tuna z ribiško palico zaposliti tri mavretanske
00002 CONO IN IZSTOP IZ NJEGA 1 . Razen plovil za **ribolov** tuna in plovil s površinskimi parangali
00002 Skupnosti , ki je imetnik dovoljenja za **ribolov** v ribolovni coni Mavretanije , razen plovil
00002 prizadevanju , da se prepreči nezakoniti **ribolov** v ribolovni coni Mavretanije , ki ogroža
00004 OBRAZLOŽITVENI MEMORANDUM Dostop plovil ES za **ribolov** tunov na ribolovna območja v osrednjem
00004 Določa pogoje dostopa evropskih plovil za **ribolov** tuna v vode FDM in okvir za prispevke ES
00004 možnosti ribolova ES ima dovoljenje za **ribolov** 6 plovil z zapornimi plavaricami in 12

Besedila z drugih virov

- pretvorba iz .doc, .rtf, .html v .txt: doc2txt (http://doc2txt.com/index.php), $29.95
- pretvorba iz .pdf v .txt: AAA (http://www.aaapdf.com/pdf2text.htm), $29.50

Označevanje besedil

- spletna lematizacija in oblikoskladenjsko označevanje (RDR in CLOG) - http://nl2.ijs.si/analyze/
- poleg slovenščine še madžarščina, estonščina, češčina, romunščina, angleščina
- različni izhodni formati (tudi TEI XML)

Korpusni pregledovalniki

- tip in zapis korpusa?
- velikost?
- eno- ali večjezični?
- uporabniške zahteve (načini poizvedb, statistične obdelave)?
- ...

Nekaj orodij

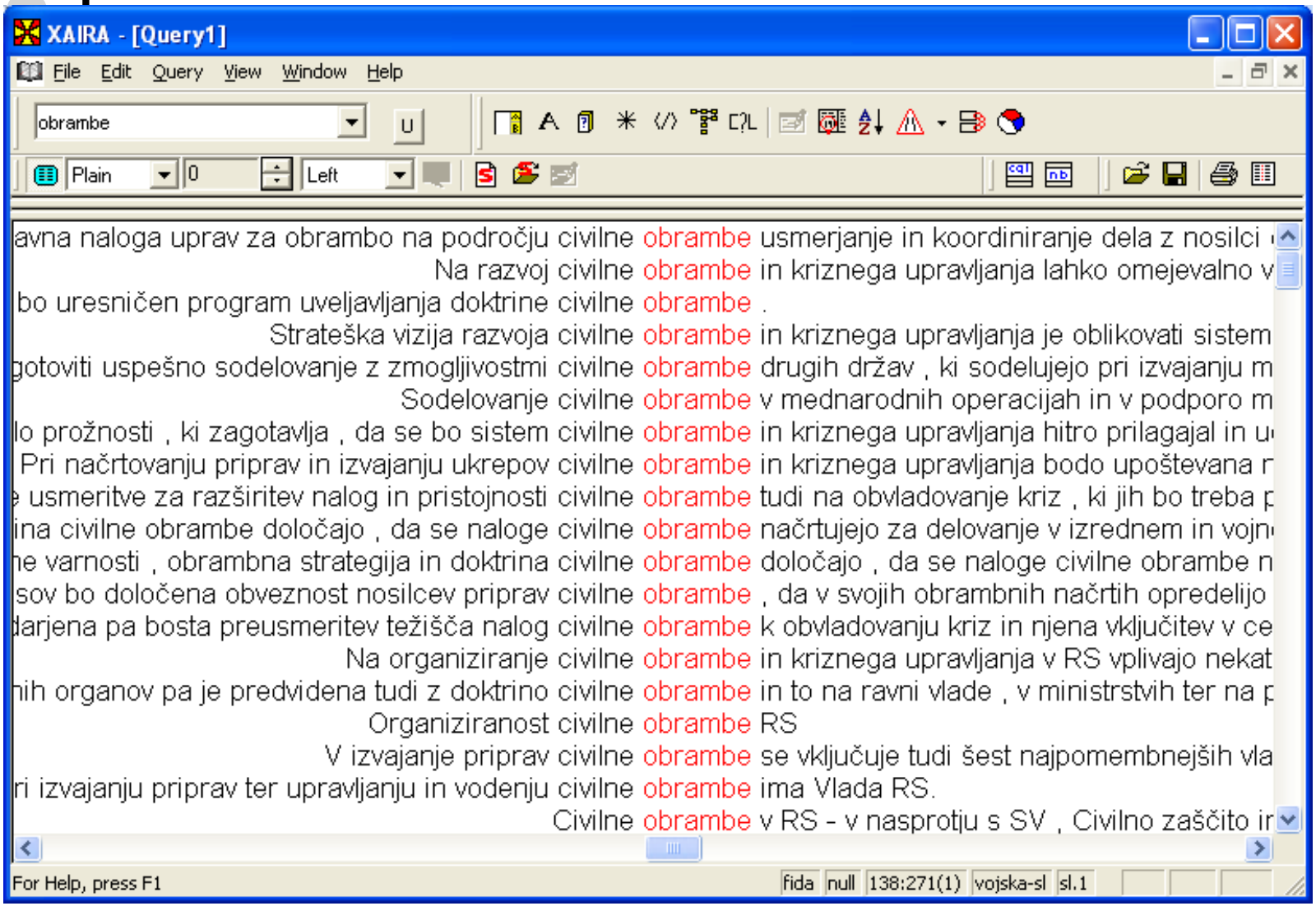
- **WordSmith Tools, avtor Mike Scott, vključuje orodja za besedne sezname, konkordance, analizo ključnih besed itd., distribucija prek Oxford University Press**
- **Concordance - besedni seznami, konkordance, avtor R.J.C. Watt**
- **Monoconc - konkordančnik, avtor M. Barlow, distribucija prek podjetja Athelstan**
- **Paraconc - konkordančnik za vzporedne korpuse, avtor M. Barlow**
- **Multiconcord - konkordančnik za vzporedne korpuse, avtor David Wools**
- **LEXA *Corpus Processing Software,* distribuira ICAME**
- **Xaira – bivša SARA, korpusni pregledovalnik, podpira TEI XML (prosto dostopna)**

WordSmith Tools

- primeren za obdelavo enojezičnih korpusov
- ni primeren za označene korpuse
- le primitivne poizvedbe
- besedila morajo biti pretvorjena v golo besedilo (TXT)
- podpira različna kodiranja (ANSI, Unicode)
- omogoča shranjevanje rezultatov v treh oblikah (.lst, .txt, .xls)
- na voljo tudi v slovenščini
- stane £59.50

Xaira

- zastonj
- podpira XML in TEI
- podpira poizvedbe z nadomestnimi znaki, CQL, XML, regularne izraze
- vključuje izračun kolokatorjev
- vključuje možnost dodajanja lastnih označevalnih ravni
- nima možnosti izdelave seznamov


XAIRA - [Query1]
File Edit Query View Window Help
obrambe
U
Plain
0
Left
avna naloga uprav za obrambo na področju civilne obrambe usmerjanje in koordiniranje dela z nosilci
Na razvoj civilne obrambe in kriznega upravljanja lahko omejevalno v
bo uresničen program uveljavljanja doktrine civilne obrambe .
Strateška vizija razvoja civilne obrambe in kriznega upravljanja je oblikovati sistem
gotoviti uspešno sodelovanje z zmogljivostmi civilne obrambe drugih držav , ki sodelujejo pri izvajanju m
Sodelovanje civilne obrambe v mednarodnih operacijah in v podporo m
lo prožnosti , ki zagotavlja , da se bo sistem civilne obrambe in kriznega upravljanja hitro prilagajal in u
Pri načrtovanju priprav in izvajanju ukrepov civilne obrambe in kriznega upravljanja bodo upoštevana r
usmeritve za razširitev nalog in pristojnosti civilne obrambe tudi na obvladovanje kriz , ki jih bo treba p
ina civilne obrambe določajo , da se naloge civilne obrambe načrtujejo za delovanje v izrednem in vojn
he varnosti , obrambna strategija in doktrina civilne obrambe določajo , da se naloge civilne obrambe n
sov bo določena obveznost nosilcev priprav civilne obrambe , da v svojih obrambnih načrtih opredelijo
darjena pa bosta preusmeritev težišča nalog civilne obrambe k obvladovanju kriz in njena vključitev v ce
Na organiziranje civilne obrambe in kriznega upravljanja v RS vplivajo nekat
nih organov pa je predvidena tudi z doktrino civilne obrambe in to na ravni vlade , v ministrstvih ter na p
Organiziranost civilne obrambe RS
V izvajanje priprav civilne obrambe se vključuje tudi šest najpomembnejših vla
ri izvajanju priprav ter upravljanju in vodenju civilne obrambe ima Vlada RS.
Civilne obrambe v RS - v nasprotju s SV , Civilno zaščito ir
For Help, press F1
fida null 138:271(1) vojska-sl sl.1

ParaConc

- pregledovalnik vzporednih korpusov
- vsebuje tudi orodje za vzporejanje (poravnavo)
- ne podpira označenih besedilnih formatov
- vključuje izračun kolokatorjev
- podpira regularne izraze
- stane $95

Zaključne pripombe

- kaj nam še manjka?
  - slovar sodobnega slovenskega jezika
- aplikacije?
  - strojno prevajanje
  - semantične aplikacije
  - …
- viri in JT za jezikoslovce
  - prosto dostopna orodja za vse ravni označevanja
    - oblikoskladenjsko
    - skladenjsko
    - semantično
- korpus ≠ corpus delicti