

Besede v besedilih

Univerza v Ljubljani
Filozofska fakulteta
Oddelek za prevajalstvo

Leksika slovenskega
jezika

Študijsko leto 2009/10



asist. dr. Darja Fišer

Besedišče v besedilih

- distribucija pogostih & redkih besed v korpusu
 - slovnične vs. polnopomenske besede
 - pogosto vs. redko besedišče
- povezanost najpomembnejših kolokatov (npr. opravi, pasti, rukniti izpit) & jeder (npr. vprašanje - odgovor)
- besede, ki se pojavljajo blizu skupaj, si delijo semantične lastnosti > kohezija
- kohezija = stare + nove info
 - formulaične, rutinizirane, institucionalizirane fraze
 - variantnost (npr. zvezda je rojena, rojstvo zvezde)
 - razmerje med pojavnicami in različnicami
 - določanje stopnje težavnosti, pripisovanje avtorstva, ugotavljanje reprezentativnosti korpusa

JOS - lematizacija & označevanje

The screenshot shows a web browser window with the address bar containing the URL http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika, which is circled in red. The browser's tab bar shows several tabs, including 'Word List Form', 'ToTaLe analyser', and 'Wikivir:Slovenska leposlovna klas...'. The main content area features a Wikivir logo on the left and a navigation menu. The main heading is 'Wikivir:Slovenska leposlovna klasika'. Below the heading, there is a paragraph of text starting with 'Seznam za digitalizacijo in postavitve na Wikivir obsega avtorje in naslove, ki jih druge zbirke...'. A red arrow points from the text 'išči pod "Dokončano"' to the main heading.

Wikivir:Slovenska leposlovna klasika - Wikivir

[http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika](#)

svasishth rlevy vis masaža blumau Navodila za ...acijo strani aurora truffles2 truffles1 cookies tuffles ny magazine ambasadorji

Word List Form ToTaLe analyser Wikivir:Slovenska leposlovna klas...

Poskusite E

[projektna stran](#) [pogovor](#) [uredite stran](#) [zgodovina strani](#)

Novost: [Popolni pomočnik](#) za dodajanje predlog! Preizkusite

Wikivir:Slovenska leposlovna klasika

Seznam za digitalizacijo in postavitve na Wikivir obsega avtorje in naslove, ki jih druge zbirke (Luinova Beseda, Hladnikova Zbirka slovenskih leposlovnih besedil, Jakopinova Nova beseda, Digitalna knjižnica Slovenije, Slovenska literatura v Intratext Digital Library) še ne vsebujejo in spadajo med kanonizirano in trivialno klasiko. V nadaljevanju pa pride na vrsto tudi objava že digitaliziranih besedil, ki se v teh zbirkah nahajajo v drugih besedilnih formatih ali pa jih je treba peljati še skozi korekturo. K sodelovanju so vabljeni zlasti študentje slovenščine in književnosti (najbolj zavzete bomo priporočili profesorjem za višjo oceno), sicer pa je dobrodošel vsakdo. Študentje med počitnicami s popraviljanjem lahko tudi **zaslužijo**. Za besedila avtorjev, ki še žive ali so umrli po letu 1939, je treba pridobiti dovoljenje lastnikov avtorskih pravic, zato so taki na seznamu redkejši. Zainteresirani avtorji, javite se sami! Zainteresirani wikivirovci, povprašajte jih sami!

WIKIVIR

navigacija

- Glavna stran
- Portal občestva
- Pod lipo
- Izbrana besedila
- Zadnje spremembe
- Naključno besedilo
- Pomoč
- Denarni prispevki

print/export

JOS - lematizacija & označevanje

The screenshot shows a web browser window titled "ToTaLe analy". The address bar contains the URL "http://nl.ijs.si/jos/analyse/" which is circled in red. Below the address bar, there are several tabs: "svasishth", "rlevy", "vis", "masaža", "blumau", "Navodila za ...acijo strani", "aurora", "truffles2", and "truffles1". The main content area has a blue background and contains the following elements:

- A text input area with the placeholder text: "There is no pre-set limit for the size of the text, however, the practical limit (server timeout)".
- A section titled "Type or paste in text:" with a text area containing the Slovenian text: "Tistega večera sem preveč popil, zgodilo se je mesec dni po tem, ko sem izvedel, da me žena vara. Dogodek v Ankaranu je bila dramatična nesreča. Dekle je ob vzvratni vožnji začelo vpiti, da bi jo utišal, sem prijel nož. Prišlo je do prerivanja in umrla je."
- A section titled "Plain text file in UTF-8:" with a "Choose File" button circled in red and the text "no file selected".
- A section titled "Analyse the text and" followed by a "show" button, the word "or", a "download" button circled in red, and the text "the results."

Sketch Engine - Corpus Builder



Sketch Engine

user: Darja Fišer

[Help](#) [Change passwd](#) [WebBootCaT](#) [CorpusBuilder](#) [Bug reporting](#) [News](#) [Logout](#)

New beta

On beta (<http://beta.sketchengine.co.uk/>) we are testing our new infrastructure, 'Corpus Architect'. Please do try it! Note that

- * accounts are separate: if your account does not work on beta, set up a new free-trial one
- * some specialist corpora are not yet available on Corpus Architect
- * WebBootCaT functionality is available on the new system, but not in the same way. To build a WebBootCat corpus in Corpus Architect, you first need to click the button for 'build new corpus' and follow steps from there.

Sketch Engine - Corpus Builder

Corpus Builder

[homepage](#) | [corpora](#)

Note

We have now started the process of upgrading the service to a new version of the software and new servers. accounts (same username and password) set up on the new system. Trial users can set up a new trial account

If you want to transfer a corpus to the new system we recommend that you download the corpus and re-upload it with the same corpus name, and we shall do our best to assist you.

Corpus list

Create new corpus [from template](#)

Corpus ID	Name	Metadata	Structures	Tokens [?]	
bla	bla	title	p, s, g, doc	56	[open in SkE] [expert config] [access] [edit] [del]

Sketch Engine - Corpus Builder

Create new corpus from template

Corpus ID [?]	<input type="text" value="kratica"/>
Corpus name [?]	<input type="text" value="ime korpusa"/>
	<p><input type="radio"/> Plain Upload plaintext files in any language. Converting to vertical only.</p> <p><input checked="" type="radio"/> Tagged, WS Upload POS-tagged and lemmatized texts in vertical format. Building word sketches enabled.</p> <p><input type="radio"/> English, TreeTagger Upload English texts. POS-tagger and lemmatizer (TreeTagger) enabled.</p> <p><input type="radio"/> English, TreeTagger+WS Upload English texts. POS-tagger and lemmatizer (TreeTagger) and word sketches enabled.</p>

	<p><input type="radio"/> Bulgarian, TreeTagger+WS Upload Bulgarian texts. POS-tagger and lemmatizer (TreeTagger) and word sketches enabled.</p>
Uploaded files metadata [?]	<p><input type="radio"/> ID, title, date, genre, domain, place of publication</p> <p><input type="radio"/> Title, Sub-field</p> <p><input checked="" type="radio"/> Title</p>
Uploaded files encoding [?]	<input type="text" value="UTF-8"/>

Create corpus

Reset

Sketch Engine - Corpus Builder

Create new corpus from template

Corpus **kratica** created successfully.

Initial actions OK.

[Start working with corpus](#)

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge >>> encodevert >>> compile word sketches >>> recompute scores in ws >>> compile thesaurus](#)

Files list

[Upload new file](#)

POS-tagged + lemmatized vertical	Tokens [?]	[inv]

☐ rebuild dependency files

☐ run in background

[Build selected](#) [?]

Sketch Engine - Corpus Builder

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge >>> encodevert >>> compile word sketches >>> recompute scores in ws >>> compile thesaurus](#)

New file

File [?]	<input type="button" value="Choose File"/> no file selected
File type [?]	POS-tagged + lemmatized vertical <input type="button" value="v"/>
Title	<input type="text" value="nasloj.besedila"/>

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge >>> encodevert >>> compile word sketches >>> recompute scores in ws >>> compile thesaurus](#)

New file

Uploaded file successfully saved as **text-2009-11-04T14.40.ske.vert.**

[Upload next file](#)

[Go to files list](#)

Sketch Engine - Corpus Builder

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

Merge

1 files merged successfully.

[View current merged file](#)

☒ [text-2009-11-04T14.40.ske.vert](#)

[Merge selected files](#) [\[?\]](#)

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

encodevert [\[?\]](#)

☐ run in background

[Run](#) [\[?\]](#)

[Reset](#)

Sketch Engine - Corpus Builder

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

encodevert [?]

Action OK.

[See log](#)

naloži fajl ws.txt
(najdeš na eučilnici)

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

compile word sketches [?]

Word sketch definition file [Choose File](#) No file selected

File type [text](#)

☐ run in background

[Run](#) [?] [Reset](#)

Sketch Engine - Corpus Builder

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

compile word sketches [\[?\]](#)

Action OK.

[See log](#)

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

recompute scores in ws [\[?\]](#)

Statistical function [Saliency](#)

☐ run in background

[Run](#) [\[?\]](#) [Reset](#)

Sketch Engine - Corpus Builder

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

recompute scores in ws [\[?\]](#)

Action OK.

[See log](#)

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge](#) >>> [encodevert](#) >>> [compile word sketches](#) >>> [recompute scores in ws](#) >>> [compile thesaurus](#)

compile thesaurus [\[?\]](#)

☐ run in background

Run [\[?\]](#)

Reset

Sketch Engine - Corpus Builder

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge >>> encodevert >>> compile word sketches >>> recompute scores in ws >>> compile thesaurus](#)

compile thesaurus [?]

Action OK.

[See log](#)

Corpus: kratica

[files](#) | [concordances](#) | [logs](#) | [access privileges](#)

[merge >>> encodevert >>> compile word sketches >>> recompute scores in ws >>> compile thesaurus](#)

Allowed users

Username	Full name	[inv]
No users		

Add users

Regular expression search

Username	Full name	[inv]
ljub01	Simon Krek	<input type="checkbox"/>
ljub02	Nataša Logar	<input type="checkbox"/>
ljub03	Unversity of Ljubljana, vaje	<input type="checkbox"/>

ljub49	Špela Vintar	<input checked="" type="checkbox"/>
ljub50	Marko Stabej	<input type="checkbox"/>
ljubicad	Ljubica Damevska	<input type="checkbox"/>

Sketch Engine - Corpus Builder

Corpus: kratica

files **concordances** logs | access privileges

merge >>> encodevert >>> compile word sketches >>> recompute scores in ws >>> compile thesaurus

Allowed users

Enabled access to 1 user(s)

Username	Full name	[inv]
ljub03	Unversity of Ljubljana, vaje	<input type="checkbox"/>

Disable access to selected users

Home Concordance **Word List** Word Sketch Thesaurus Sketch-Diff

Corpus: ime korpusa

Query:

Keyword(s) ☐

Home Concordance Word List **Word Sketch** Thesaurus Sketch-Diff

All words All lemmas

Word list options