

Konkordančniki CLARIN.SI

Jakob Lenardič

Inštitut za novejšo zgodovino

Predavanje NUK
2023-02-22

Kazalnik

- 1 Uvod
- 2 Trije konkordančniki CLARIN.SI
- 3 noSketch Engine (Crystal)
- 4 Zaključek

Na splošno o konkordančnikih

- **Konkordančnik:** programska oprema za iskanje po jezikovnih korpusih
- **Konkordance:**

Vprašujemo vas, kateri starostniki hodijo	nakupovat	med 22. in 6. uro? Seveda imamo možnost
ozorila pridejo ob petih zjutraj z bagerjem	popravljat	streho? Tako deprivilegiranemu delu mladir
e v petek zvečer, pa smo šli raje še malo	žurat	. Vedeli smo, da je gneča, a nimamo druge
a igralca v ramo, ga z vabilom, če se gre	tepst	, poskušal sprovocirati in malo je manjkalo,

Pogled KWIC – “KeyWord In Context”

Svetovni konkordančniki

Sketch Engine (<https://www.sketchengine.eu/>)

- Razvili Lexical Computing na Češkem
- Iskanje po več kot 700 korpusih
- Komerčni konkordančnik

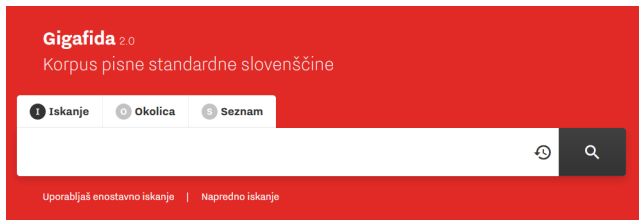
“English Corpora” (<https://www.english-corpora.org/>)

- Poglavitni konkordančnik za ameriško angleščino, npr. *Corpus of Contemporary American English (COCA)*
- Omejen prost dostop

Drugi konkordančniki za specifične jezikovne skupine
npr. *Korp* (nordijski in baltski jeziki, finščinina)

Konkordančniki v Sloveniji

- Najbolj znan je konkordančnik GigaFide
<https://viri.cjvt.si/gigafida/>



- Konkordančniki CJVT za druge korpusne: *Šolar*, *Lektor*, *GOS*
- Drugi namenski konkordančniki: *EvroTerm*

Trije konkordančniki CLARIN.SI

Kratek pregled

noSketch Engine (Bonito)

- Stara verzija nekomercialnega noSketch Engine
- Lexical Computing ga ne vzdržuje več, ampak:
- Najpopularnejši konkordančnik CLARIN.SI

noSketch Engine (Crystal) ⇐ danes podrobneje o temle

- Nova verzija nekomercialnega noSketch Engine
- Navodila za uporabo: [tukaj](#)

KonText

- Omogoča prijavo AAI (EduGain / EduRoam)
⇒ izdelava lastnih podkorpusov, hranjenje poizvedb itd.
- Navodila za uporabo: [tukaj](#)

noSketch Engine (Bonito)

`https://www.clarin.si/noske/`

NoSketch Engine

▼ 🔍

Gigafida v2.0 (referenčni, dedupliciran)

Home

Search

Word list

Corpus info

My jobs

User guide ↗

Corpus: Gigafida v2.0 (referenčni, dedupliciran) ▼

Simple query:

Make Concordance

[Query types](#) [Context](#) [Text types](#) ?

noSketch Engine (Crystal)

<https://www.clarin.si/ske>

CONCORDANCE

Gigafida v2.0 (referenčni, dedupliciran)



BASIC

ADVANCED

ABOUT

Simple search ?

abc

Text types ? ▾

SEARCH



KonText

<https://www.clarin.si/kontext/corpora/corplist>

kon text Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: [Gos 1.1.1 \(referenčni, govorni\)](#) | Query: [kriza](#) (69 hits)

Govorni posnetki

Hits: 69 | i.p.m.: 64.83 (related to the whole corpus) | ARF: 27.8 | Result is sorted 1 / 2

Line selection: simple

☐	gos005	c eee so stari svetovni problemi kot je revščina	poška	kriza	tudi že že traja in traja	+ [dobro
☐	gos005	kdo je reku da je treba zdaj rešit pa finančna		kriza	vsi oblublajo se bo boljše pa tranzicijski	
☐	gos020	?] + [to smo že rekli] + [eem] + [		kriza	v svetu ? ... torej kje se je začela tista	
☐	gos020	hotla več Amerike] + [Zedea] + [v Zedea je izbruhnila svetovna gospodarska		kriza	in se potem razširila po celem ... svetu :	
☐	gos025	zavarovanje [ime] [priimek]] + [že v lanskem letu se je nakazovala		kriza	na na podlagi česar v zadnjih mesecih l	
☐	gos032	ni več ne ker je na splošno mmm pač postaja		kriza	in tako naprej ne] + [mhm] +	

Prikaz konkordanc v govornem korpusu *Gos 1.1.1*

noSketch Engine (Crystal)

Vstopna stran – izbor korpusa



Slovenian	ELTeC-slv (100 romanov)	5,606,063 words	OPEN
Slovenian	EU DGT-UD: Slovenian	77,865,562 words	OPEN
Slovenian	FidaPLUS (stari referenčni)	600,309,637 words	OPEN
Slovenian	FILMI (filmske kritike)	769,625 words	OPEN
Slovenian	Gigafida v1.1 (referenčni)	1,146,543,854 words	OPEN
Slovenian	Gigafida v1.1 DeDup (referenčni, dedupliciran)	922,808,492 words	OPEN
Slovenian	Gigafida v2.0 (referenčni, dedupliciran)	1,109,441,592 words	OPEN
Slovenian	Gigafida v2.0 proto (referenčni, nededupliciran)	1,483,694,219 words	OPEN
Slovenian	goo300k (starejša besedila, ročno označena)	288,965 words	OPEN
Slovenian	Gos 1.1.1 (referenčni, govorni)	1,033,614 words	OPEN
Slovenian	GosVL 4.2 (govorni, VideoLectures)	178,716 words	OPEN
Slovenian	IMP (starejša besedila)	14,405,281 words	OPEN

Dostop do okrog 100 korpusov v 33 jezikih z 20 milijardami besed

O nekaj pomembnejših korpusih (1/2)

GigaFida 2.0

- Referenčni korpus – pisna standardna slovenščina
- 1.1 milijarda besed
- Besedila (predvsem tisk) 1990–2018

Korpusi družine JANES

- Spletna komunikacija (tviti, forumi, blogi)
- Pomembni za raziskavo spletnega jezika
- Posebne oznake (normalizacija)

O nekaj pomembnejših korpusih (2/2)

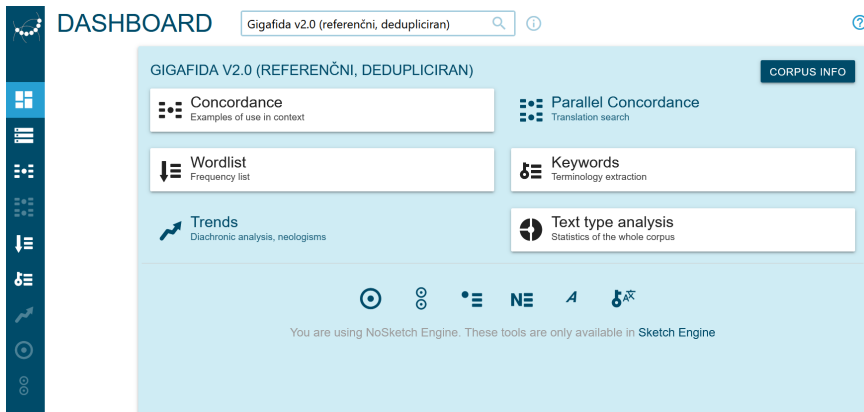
Korpusi starejše slovenščine IMP

- Razpon: 16. stoletje – 1918
- Verske knjige, *Kmetijske in rokodelske novice* (NUK!), leposlovje itd.
- Posebne oznake (normalizacija)

Korpus akademske slovenščine (KAS)

- Diplomске naloge, magistrska dela, doktorske disertacije
- Pomemben vir za preučevanje akademske slovenščine
- Označeni termini

noSke nadzorna plošča



DASHBOARD Gigafida v2.0 (referenčni, dedupliciran) ⓘ

GIGAFIDA V2.0 (REFERENČNI, DEDUPLICIRAN) **CORPUS INFO**

- Concordance**
Examples of use in context
- Parallel Concordance**
Translation search
- Wordlist**
Frequency list
- Keywords**
Terminology extraction
- Trends**
Diachronic analysis, neologisms
- Text type analysis**
Statistics of the whole corpus

⦿ ⦿ ⋮ NE A 🔑^{AK}

You are using NoSketch Engine. These tools are only available in Sketch Engine

- Konkordance
- Besedni seznam
- Ključne besede
- Analiza besedilnih vrst

Pomembno: CORPUS INFO

Korpusna statistika

TEXT TYPE ANALYSIS

Gigafida v2.0 (referenčni, dedupliciran)



Structures and text types

text - author

text - class

text - date

text - id

text - publisher

text - source

text - title

Show

Token coverage

Subcorpus

none (the whole corpus)

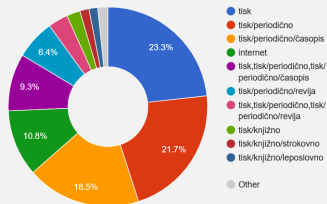
Filter results



Items: 15, Total frequency: 6,873,401,681



text - class



Konkordance 📊 (1/3)

CONCORDANCE Gigafida v2.0 (referenčni, dedupliciran) ⓘ

BASIC **ADVANCED** ABOUT

Query type ⓘ

- simple
- lemma**
- phrase
- word
- character
- CQL

Part of speech

- any**
- samostalnik
- glagol
- pridevnik
- prislov
- zaimék
- predlog
- veznik

Lemma

volivec

✓ A = a?

- Vrste iskanja: *simple*, *lemma*, *CQL* itd.
- CQL: glej **User Manual** ali **YouTube**

Konkordance 🗃️ (2/3)

The screenshot shows the GigaFide concordance interface. On the left is a dark blue sidebar with various icons. The main area has a light blue header with 'Subcorpus ②' and a dropdown menu showing 'none (the whole co ...)'. Below this is 'Filter context ②' with a dropdown arrow. Underneath is 'Text types ② ^' with an expand/collapse icon. To the right of this are links for 'expand all' and 'collapse all'. The main content area displays three filter panels:

- text.class ^**: A list of categories including 'internet', 'tisk', and several sub-categories under 'tisk' like 'drugo', 'knjižno', 'leposlovno', 'strokovno', 'periodično', 'časopis', and 'revija'.
- text.date v**: A panel for filtering by date.
- text.year ^**: A list of years from 1990 to 1998, with a 'GO' button at the bottom right.

- Omejevanje iskalnega niza na podmnožice v korpusu
- V primeru GigaFide: leto, besedilna vrsta, avtor ipd.

Konkordance 📊 (3/3)

- Iskalni nizi se dajo obogatiti s ti. regularnimi izrazi
- **Pravopisno vprašanje:** Ali se piše *volilec* ali *volivec*?
- **Rešitev:** Poiščemo lemo **voli|lv]ec** (alternativno: **volilec|volivec**)



CONCORDANCE

Gigafida v2.0 (referenčni, dedupliciran) 🔍 ⓘ

lemma **voli|lv]ec** • 84,849
63.64 per million tokens • 0.0064% ⓘ

           **KWIC** ⓘ

☐ Details ☐ Left context ☐ KWIC ☐ Right context

1	☐	ⓘ GF0000290 • Več... uje evroskepticizem, zaradi katerega so volivci na referendumu zavrnili sporazum iz Nic	
2	☐	ⓘ GF0000290 • Več... stranko, ki bo jo zdaj volilo 21 odstotkov volivcev . Generalni sekretar ZN Kofi Annan se je	
3	☐	ⓘ GF0000866 • Dne... šno priprla, vprašanje, ki ga bo zastavila volilcem na referendumu o evropski ustavni pogo	
4	☐	ⓘ GF0000866 • Dne... i6. Hkrati se je dobrikala evroskeptičnim volilcem s trditvijo, da bi konservativna vlada razp	
5	☐	ⓘ GF0000866 • Dne... pska ustava skrajno nepriljubljena med volilci in poslovneži. Neodvisna stranka Združe	

Frekvenčni seznami ≡ (1/3)

The screenshot shows the CONCORDANCE Gigafida v2.0 interface. At the top, the title 'CONCORDANCE' is followed by a search bar containing 'Gigafida v2.0 (referenčni, dedupliciran)'. Below this, a lemma filter is set to 'voli|v|jec' with 84,849 occurrences and a frequency of 63.64 per million tokens. The main section is titled 'FREQUENCY' and has three tabs: 'BASIC', 'ADVANCED', and 'ABOUT'. The 'BASIC' tab is active, showing a table with columns for 'First word to the left', 'KWIC', and 'First word to the right'. Each column contains four buttons: 'WORD FORMS', 'PART OF SPEECH', 'TAGS', and 'LEMMAS'. To the right of the table, there is a 'More presets' section with two buttons: 'TEXT TYPES' and 'LINE DETAILS'. The interface includes a sidebar on the left with various navigation icons and a top bar with additional search and filter options.

- Sortiranje po lastnostih pojavnice (*word form* vs. *tags* vs. *lemmas*)
- Sortiranje po KWIC ali levi/desni besedi

Frekvenčni seznami ☰ (2/3)



CONCORDANCE

Gigafida v2.0 (referenčni, dedupliciran)



lemma volilec|volivec • 84,849

63.64 per million tokens • 0.0064%



...



Frequency

CHANGE CRITERIA

BACK TO CONCORDANCE



Show relative frequency



Show percentage of concordance lines

(3 items, 84,849 total frequency)

		Lemma	Frequency	Relative ?		
1	☐	volivec	58,041	43.53		...
2	☐	volilec	26,806	20.10		...
3	☐	Volilec	2	< 0.01		...

Frekvenčni seznami ≡ (3/3)

S CHANGE CRITERIA lahko naredimo frekvenčne sezname znotraj podmnožic v korpusu



CONCORDANCE

Gigafida v2.0 (referenčni, dedupliciran) ⓘ

lemma volilec/volivec • 84,849
63.64 per million tokens • 0.0064% ⓘ ...

Frequency

CHANGE CRITERIA BACK TO CONCORDANCE

☐ Show relative in text types ☐ Show relative density

(9 items, 190,777 total frequency)

	Text.class	Frequency	
1	☐ tisk	53,192	...
2	☐ tisk/periodično	51,976	...
3	☐ tisk/periodično /časopis	45,562	...
4	☐ internet	31,657	...
5	☐ tisk/periodično /revija	6,414	...
6	☐ tisk/knjižno	760	...
7	☐ tisk/knjižno /strokovno	594	...
8	☐ tisk/drugo	456	...
9	☐ tisk/knjižno /leposlovno	166	...

Kolokacije

		Lemma (lowercase)	Cooccurrences [?]	Candidates [?]	T-score	MI	LogDice ↓	
1		finančen	19,611	348,496	139.69	8.66	10.23	...
2		gospodarski	18,868	363,244	136.99	8.54	10.14	...
3		begunski	6,247	20,445	79.00	11.10	9.96	...
4		reševanje	4,837	105,686	69.34	8.36	9.09	...
5		izhod	3,525	34,160	59.29	9.53	9.04	...
6		hud	5,678	264,672	74.86	7.27	8.69	...
7		dolžniški	2,263	6,245	47.55	11.35	8.59	...
8		ukrajinski	2,276	29,834	47.62	9.10	8.44	...

Kolokacije za lemo *kriza*

Paralelne konkordance

PARALLEL CONCORDANCE

TRANS5: slovensko

simple **gladina** • 68

42.66 per million tokens • 0.0043%



align ▾



TRANS5: angleško

① JRC ECDC-TM (2012)	dvig morske gladine .	sea-level rise.
① 1984 (1983)	Nenadoma, kot kos potopljene razbitine, ki predre vodno gladino , mu je vdrla v glavo misel: "To se v resnici ne zgodi. "	Suddenly, like a lump of submerged wreckage breaking the surface of water, the thought burst into his mind: "It doesn't really happen. "
① Paleocenske plasti Liburn...	V zaporedju plasti se zrcali pozno paleocenska morska transgresija oz. dvig morske gladine .	In the succession the Late Paleocene transgression or the sea level change is reflected.
① Paleocenske plasti Liburn...	Dobro je definirana poznopaleocenska, thanetijska transgresija oz. dvig morske gladine , ki je po 5 milijonih let prekrila celotni prostor sedanje SW Slovenije (tab.1,3,4,5,7; sl.3,10).	On the top of this succession, well defined is the Late Paleocene Thanetian transgression resp. rise of the sea level that covered after 5 million years the entire region of present SW Slovenia (Pls.1,3,4,5,7; Figs.3,10).
① Paleocenske plasti Liburn...	V poznem paleocenu je morska transgresija oziroma globalen dvig morske gladine prekril platformo južnozahodne Slovenije.	The Late Paleocene transgression, or the global sea level-rise, covered platform in SW Slovenia.
① Formacijska geološka kart...	Zanimiv prispevek k poznavanju globalnih oscilacij morske gladine v zgornji kredi so podali Summesberger in sodelavci (1996a).	An interesting contribution to the knowledge of global oscillations of the sea level in the Upper Cretaceous was presented by Summesberger and others (1996a).

Lema **gladina**: slovenščina → angleščina

Oznake v drugih korpusih – *Janes-Norm*

- Korpus **Janes-Norm** ima ročno normalizacijo
- Konkordance za lemo **jaz**

	☐ Details	Left context	KWIC	Right context
1	☐ ⓘ tweet • T3 • L3	@spirulinka9 @tretjeoko pri ex sem si že @spirulinka9 @tretjeoko pri ex sem si že	jaz jaz	tud marsikaj lahko predstavljala.. tudi marsikaj lahko predstavljala ..
2	☐ ⓘ forum • T1 • L3	pa lova na kite in ostale nedolžne živali ! pa lova na kite in ostale nedolžne živali !	Jaz jaz	sploh ne stavim na prošnje. Jaz g sploh ne stavim na prošnje . jaz ě
3	☐ ⓘ forum • T1 • L3	e živali ! Jaz sploh ne stavim na prošnje. živali ! jaz sploh ne stavim na prošnje .	Jaz jaz	grem direkt do delodajalca oz. na grem direkt do delodajalca oz. na
4	☐ ⓘ tweet • T3 • L3	!....[::]))) @iNinaromsek mela saaansooo, :::))) @ininaromsek imela šanso ,	js jaz	grem pa kr ze dons :P se prevoz grem pa kar že danes :p še prevoz
5	☐ ⓘ tweet • T3 • L3	udn zgleda. @nejcjemec to sam pletes? idno zgleda . @nejcjemec to sam pleteš ?	js jaz	bi tut rabil btw. vem pa kako iz pl bi tudi rabil btw . vem pa kako iz pl

Oznake v drugih korpusih – IMP

- Korpusi starejše slovenščine **IMP** tudi normalizirani
- Konkordance za lemo *jaz*

☐ Details	Left context	KWIC	Right context
1 ☐ ⓘ Biblija (vzorec... u odtrefsite prah od valhjih nug : Sa rifnizo i odtresite prah od vaših nog : za resnico	jelt jaz	vam povém, de téh Sodomiterjeu inu G vam povem , da teh sodomiterjev in g	
2 ☐ ⓘ Biblija (vzorec... , na fodni dan , kakòr takimu Méftu . Pole , , na sodni dan , kakor takemu mesto . pole ,	jelt jaz	pohlem vas , kakòr Ouce , v'frédo mej V pošljem vas , kakor ovce , v sredo med vol	
3 ☐ ⓘ Biblija (vzorec... janjalitaku beshite v'enu drugu . Rifnizhnu ganjalitaku bežite v eno drugo . resnično	jelt jaz	vam povém, vy nebote Israelfka Méfta o vam povem , vi ne_boste izraelska mesta c	
4 ☐ ⓘ Biblija (vzorec... nifhtèr fkrivniga , kar bi fe nesvejdilu : Kar ništer skrivnega , kar bi se ne_zvedelo : kar	jelt jaz	vam pravim v'temmi , tu vy pravite na fvit vam pravim v temi , tu vi pravite na sve	
5 ☐ ⓘ Biblija (vzorec... òr veliku Vrabzou . Satu flejdni , kateri kuli or veliko vrabcev . zato slednji , kateri koli	mene mene	fposná pred Zhloveki , tiga hozhem jelt f spozna pred cloveki , tega hočem jaz s	
6 ☐ ⓘ Biblija (vzorec... mene fposná pred Zhloveki , tiga hozhem mene spozna pred cloveki , tega hočem	jelt jaz	fposnati pred moim Nebefkim Ozhetom : spoznati pred mojim nebeškim očetom :	

Ključne besede (1/2)



KEYWORDS

KAS (zaključna dela)  

SINGLE-WORDS 

 reference corpus: Gigafida v2.0 (referenčni, dedupliciran) (items: 6,658,361)

Word	Word	Word
1 le-ta ...	11 javno-zaseben ...	21 e-pošta ...
2 le-t ...	12 zdr-1 ...	22 e-vir ...
3 le-teh ...	13 ibid ...	23 socialno-ekonomski ...
4 vzgojno-izobraževalen ...	14 njeen ...	24 informacijsko-komunikacijski ...
5 le-to ...	15 kz-1 ...	25 povzeto ...
6 t-test ...	16 x2 ...	26 p-vrednost ...
7 le-teg ...	17 e-izobraževanje ...	27 zddpo-2 ...
8 zgd-1 ...	18 e-uprava ...	28 zp-1 ...
9 le-te ...	19 hi-kvadrat ...	29 h ...
10 e-poslovanje ...	20 zdavp-2 ...	30 katerekoli ...

Rows per page: 50  1-50 of 1,000   1

Ključne besede v **Korpusu akademske slovenščine**, ki sicer niso pogoste v GigaFidi

Ključne besede (2/2)

KEYWORDS

Gigafida v2.0 (referenčni, dedupliciran)



SINGLE-WORDS ✓



reference corpus: KAS (zaključna dela)

(items: 3,158,022)

Word	Word	Word	Word	Word
1 ugnati ...	11 kolajna ...	21 đoković ...	31 četrtkov ...	41 priigrati ...
2 oblačno ...	12 dirkač ...	22 letos ...	32 torkov ...	42 pokal ...
3 tisočak ...	13 polfinale ...	23 četrtfinale ...	33 remi ...	43 selektor ...
4 včeraj ...	14 sinoči ...	24 podprvak ...	34 koprčan ...	44 hina ...
5 derbi ...	15 lani ...	25 prvakinja ...	35 messi ...	45 hokejist ...
6 prvoligaš ...	16 favorit ...	26 reli ...	36 predlani ...	46 lakers ...
7 včerajšnji ...	17 štadion ...	27 dpa ...	37 novomeščan ...	47 borussia ...
8 sta ...	18 dončić ...	28 cibona ...	38 katanec ...	48 bečirovič ...
9 evroliga ...	19 domžalčan ...	29 menda ...	39 prvak ...	49 olimpija ...
10 drevi ...	20 afp ...	30 prevc ...	40 trumpov ...	50 ljubljankanka ...

Rows per page: 50 ▾

1–50 of 1,000



1

/ 20



Ključne besede v **GigaFidi**, ki sicer niso pogoste v Korpusu akademske slovenščine

Zaključek