Konkordančniki CLARIN.SI

Jakob Lenardič

Inštitut za novejšo zgodovino

Predavanje NUK
2023-02-22

Kazalnik

Uvod

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
ɔzorila 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
- Komercialni konkordančnik

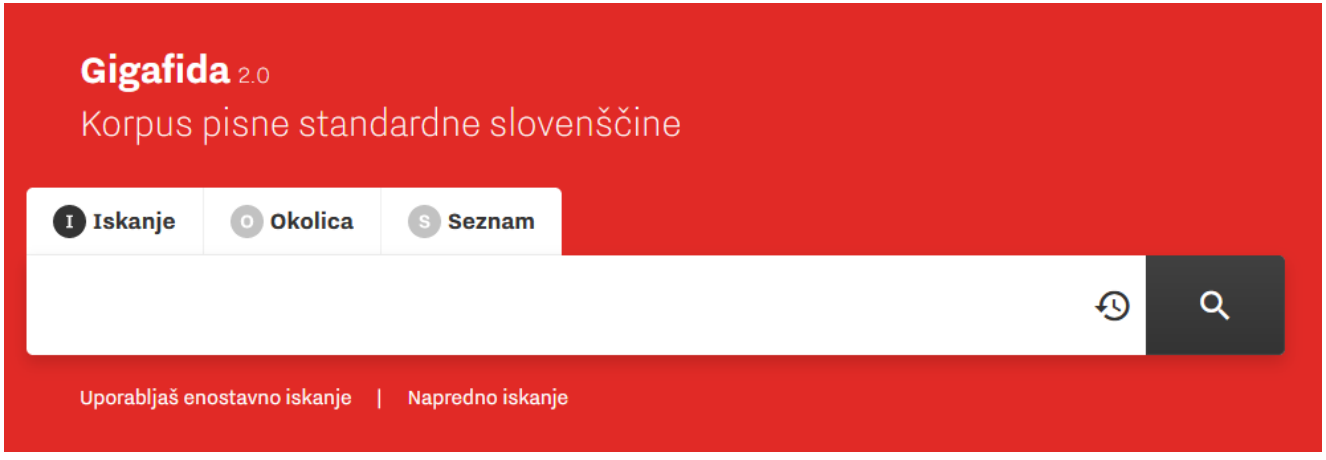
“English Corpora” (https://www.english-corpora.org/)

- Poglavnitni 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, finščina)

Konkordančniki v Sloveniji

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



- Konkordančniki CJVT za druge korpuse: *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) $\Leftarrow$ danes podrobneje o temle

- Nova verzija nekomercialnega noSketch Engine
- Navodila za uporabo: tukaj

KonText

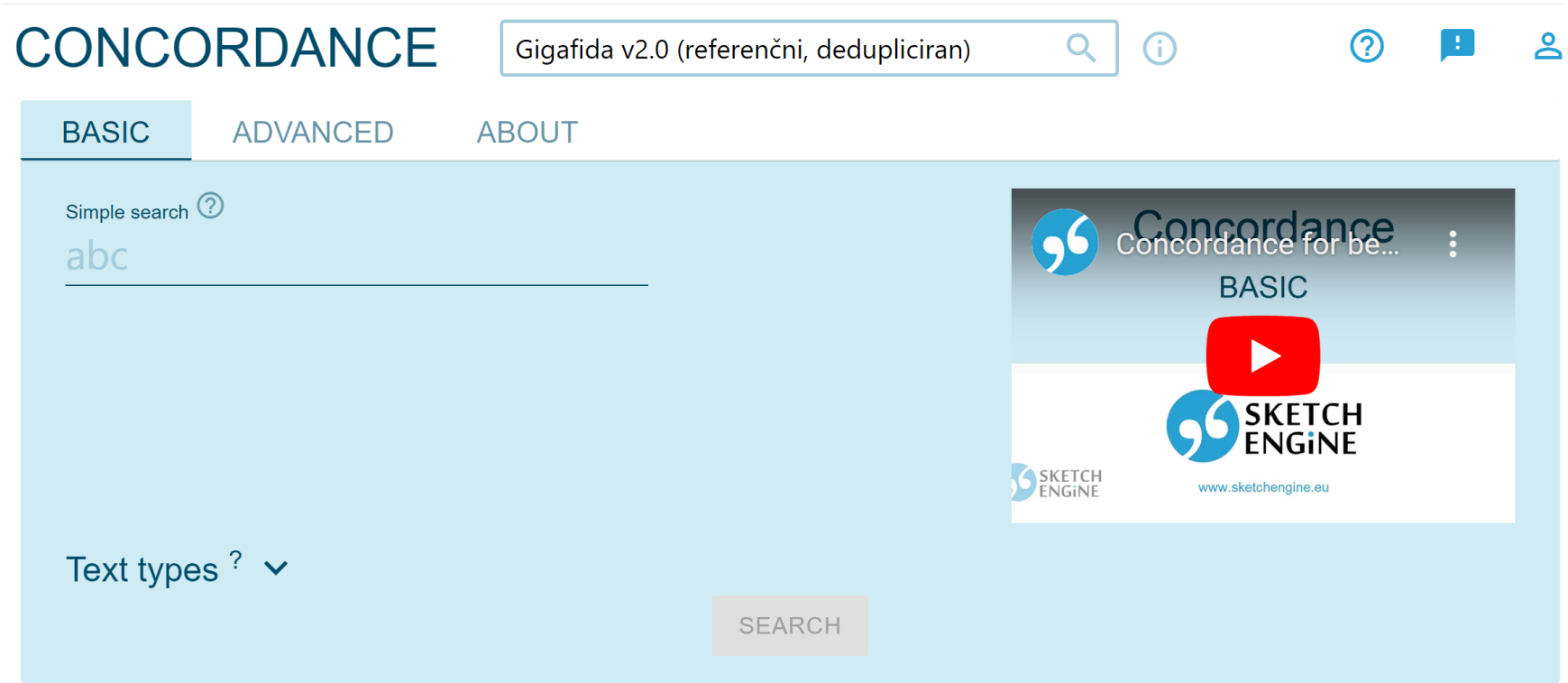
- Omogoča prijavo AAI (EduGain / EduRoam)
  $\Rightarrow$ izdelava lastnih podkorpusov, hranjenje poizvedb itd.
- Navodila za uporabo: tukaj

noSketch Engine (Bonito)

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

noSketch Engine (Crystal)

https://www.clarin.si/ske



KonText

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

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)

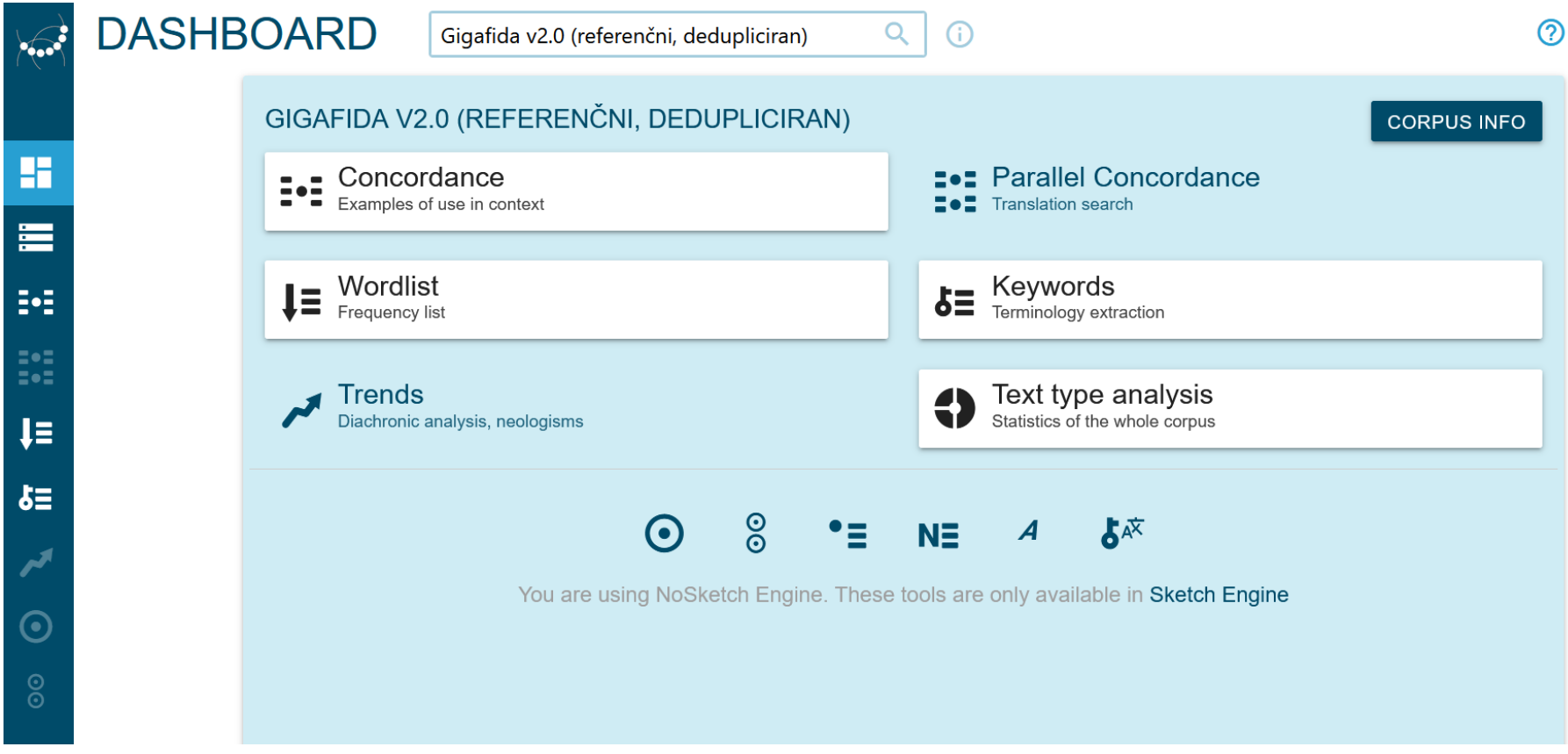
Korpusi starejše slovenščine IMP

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

Korpus akademske slovenščine (KAS)

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

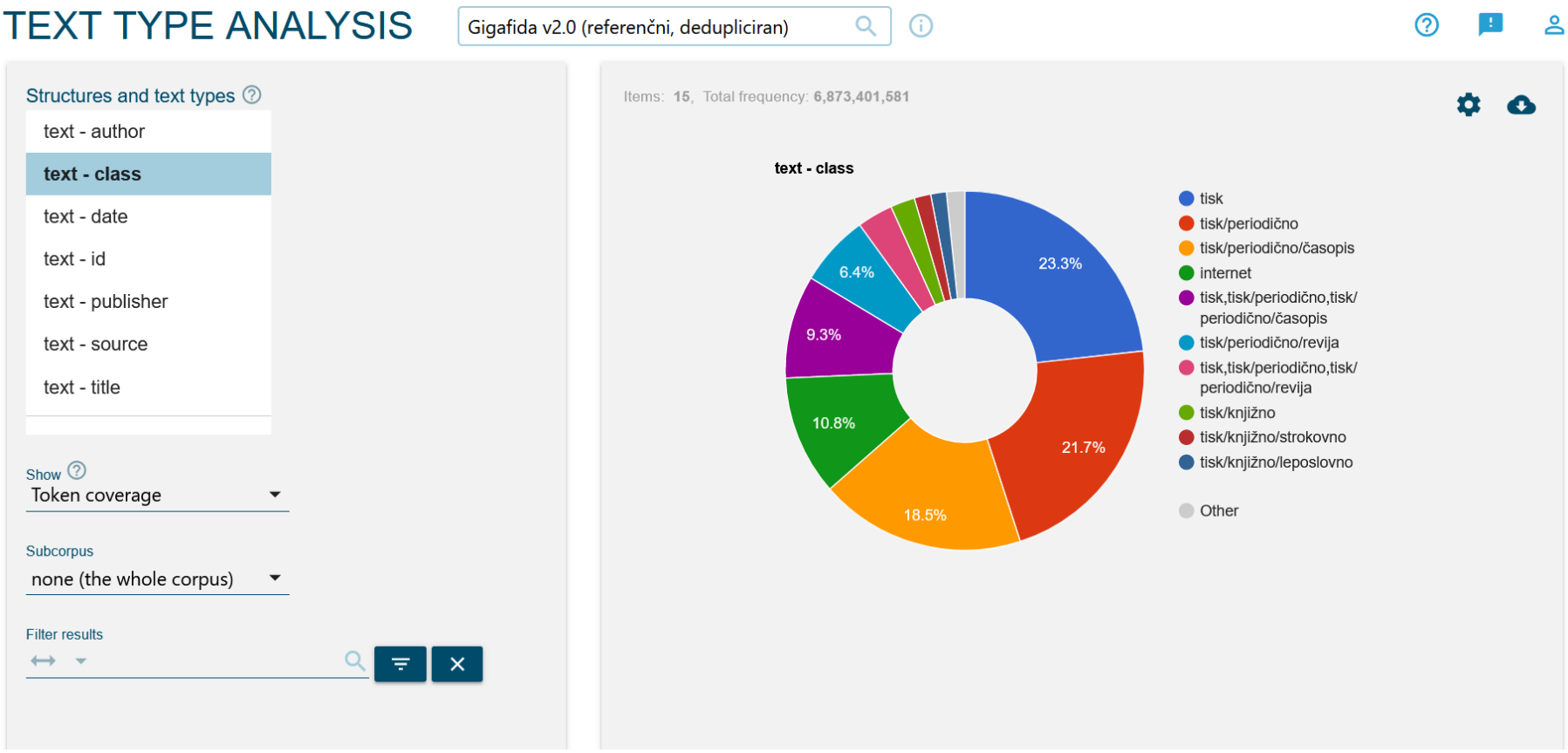
noSke nadzorna plošča



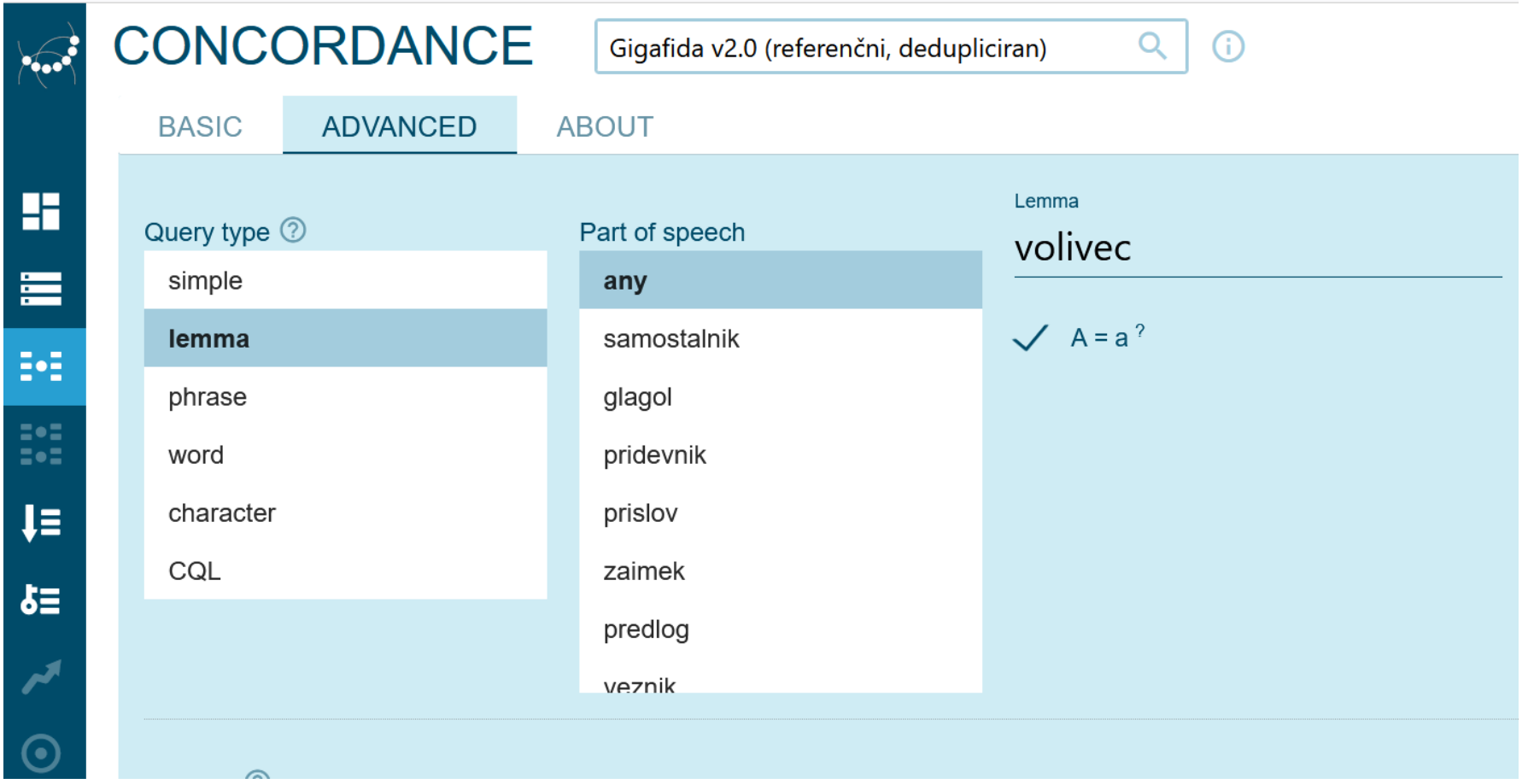
- Konkordance
- Besedni seznam
- Ključne besede
- Analiza besedilnih vrst

Pomembno: CORPUS INFO

Korpusna statistika

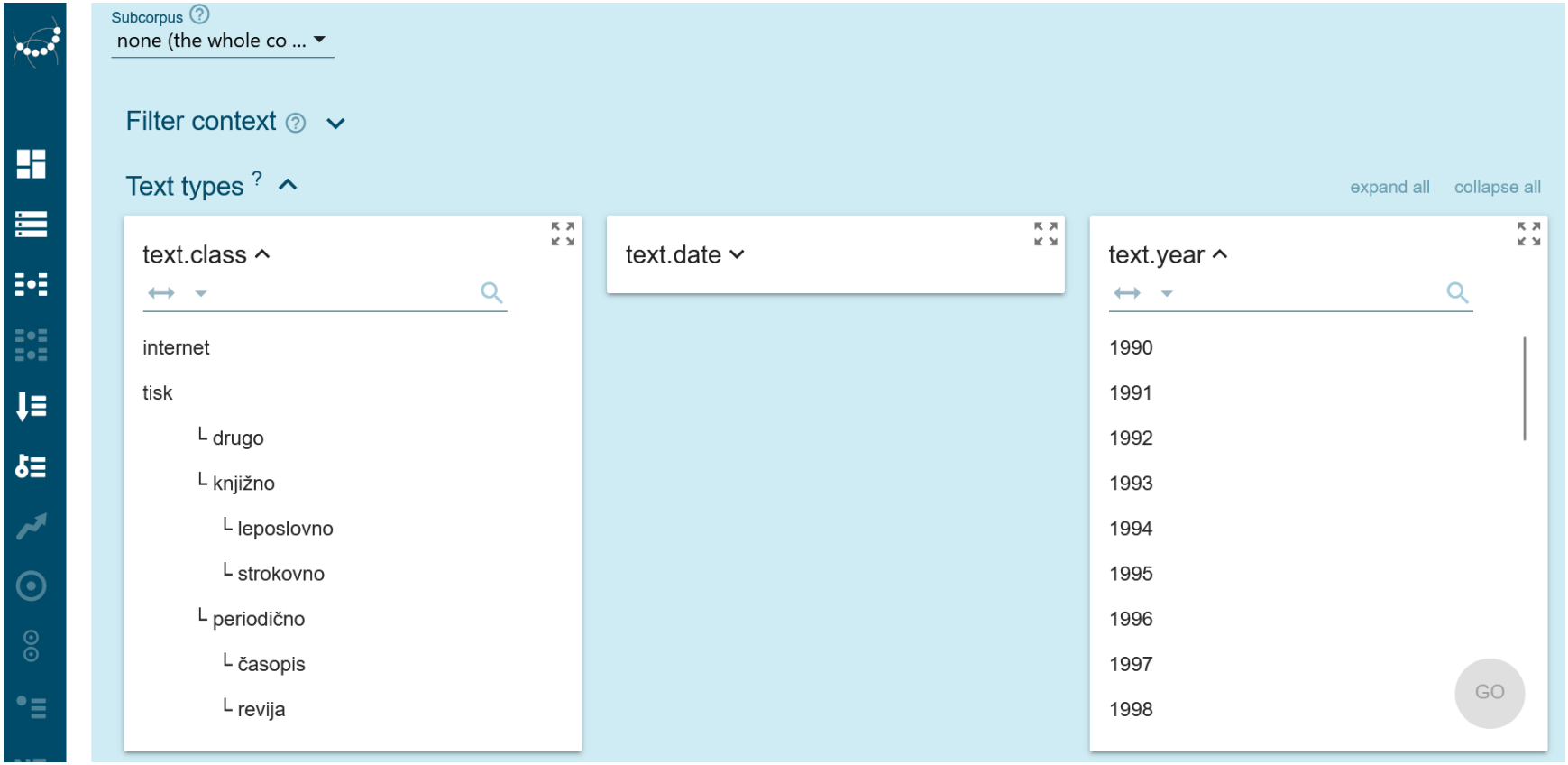


Konkordance (1/3)



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

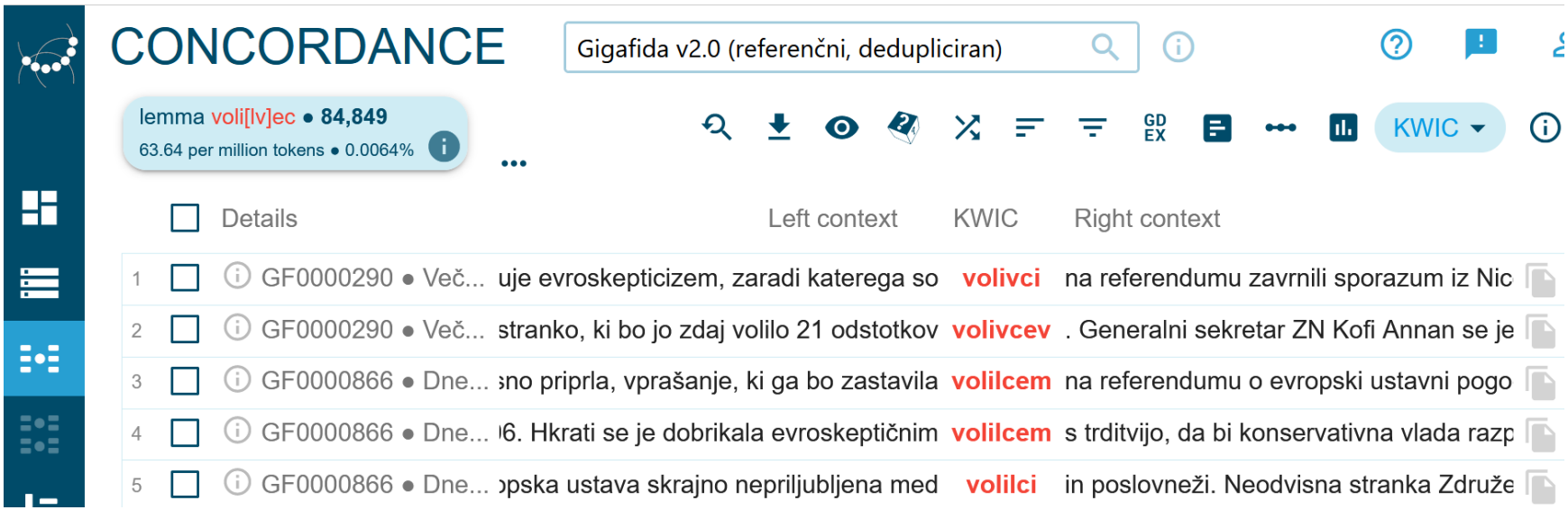
Konkordance (2/3)



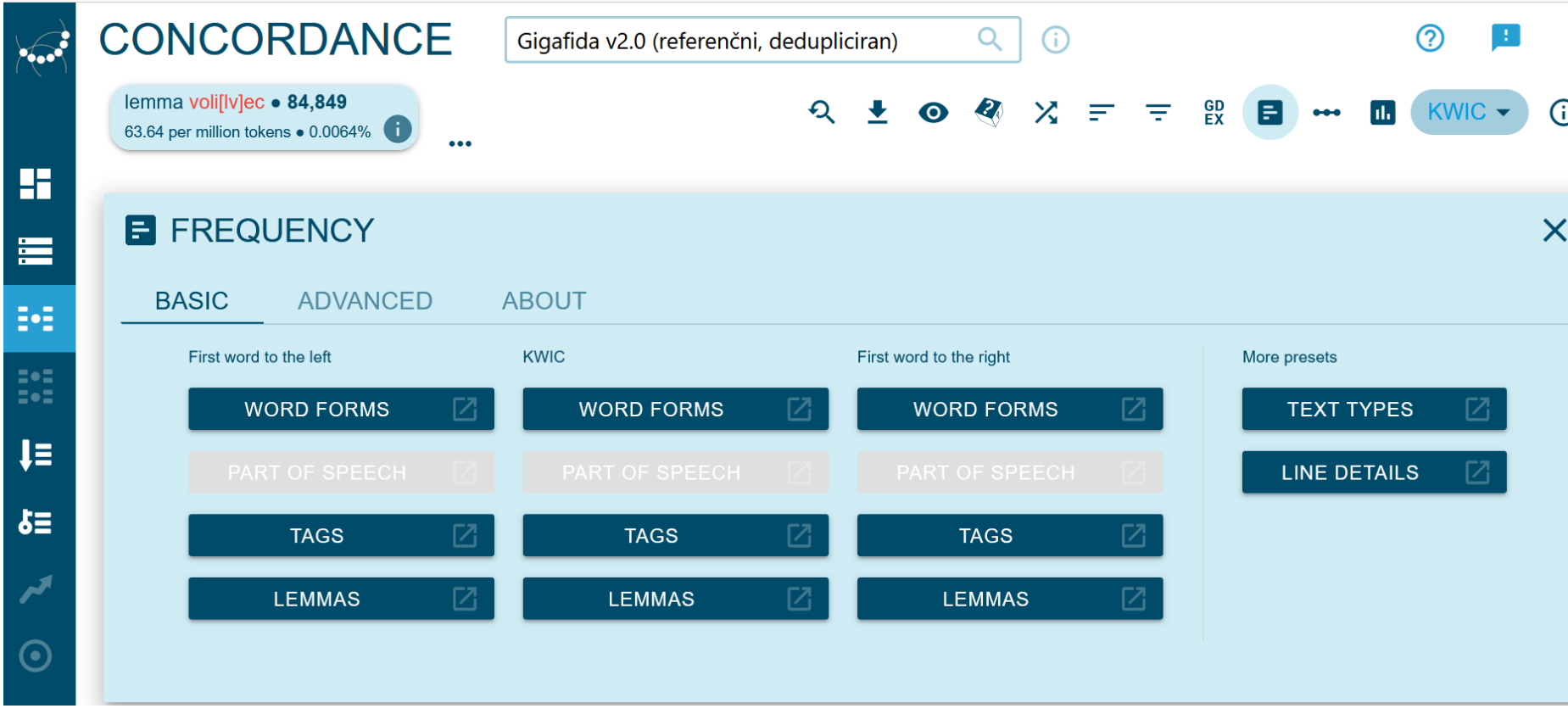
- Omejevanje iskalnega niza na podmnožice v korpusu
- V primeru GigaFide: leto, besedilna vrsta, avtor ipd.

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)

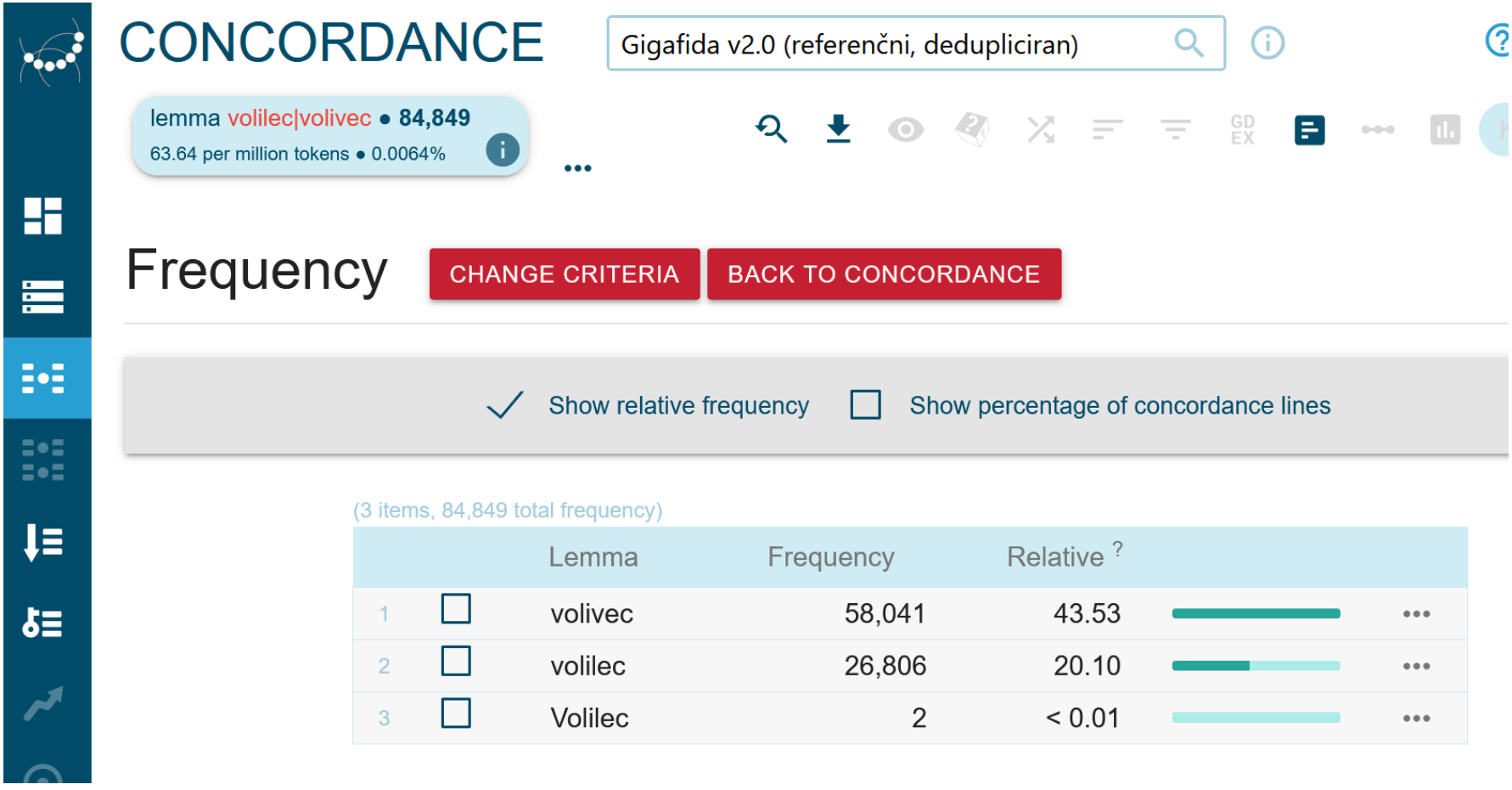


Frekvenčni seznami (1/3)



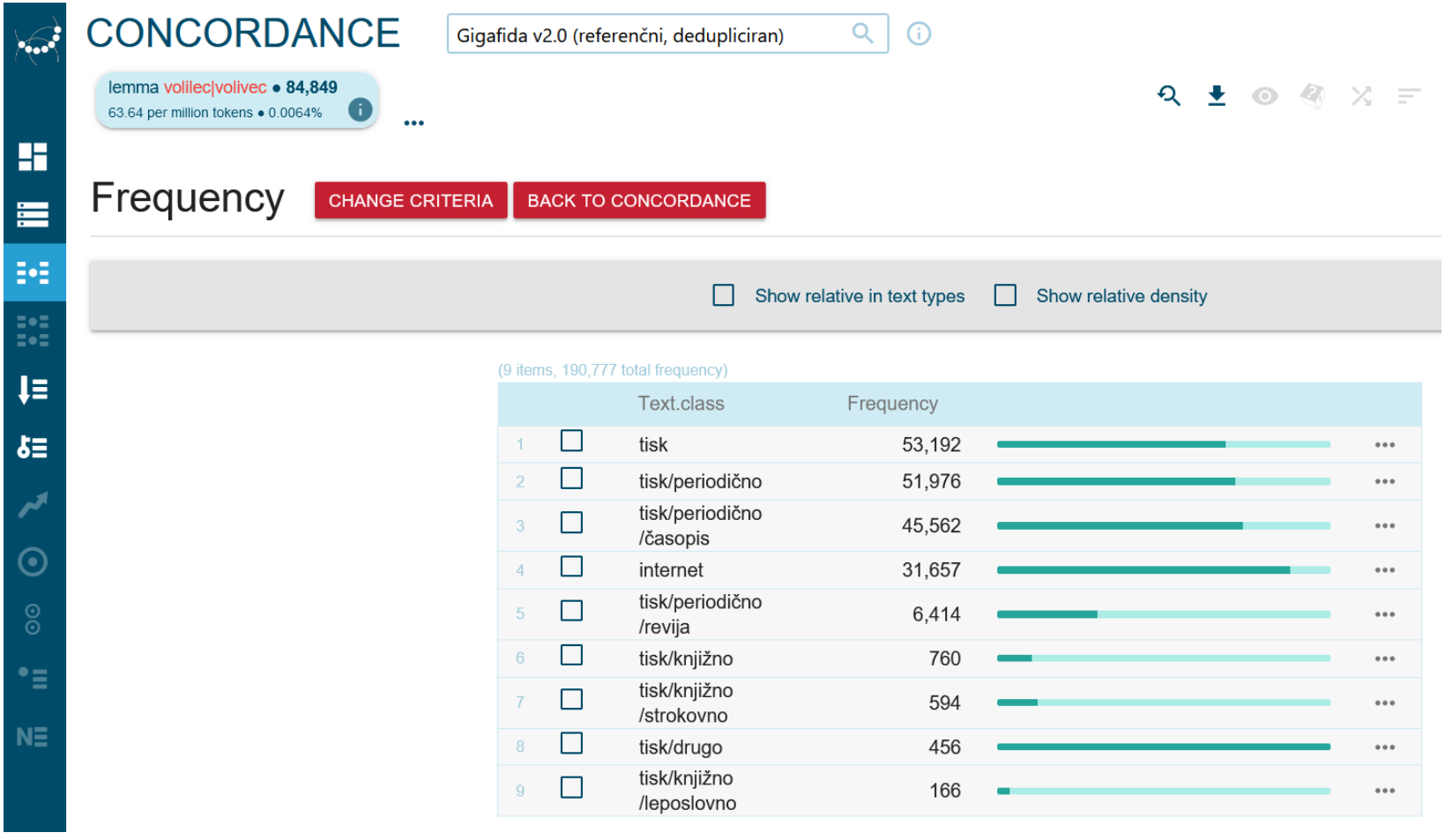
- Sortiranje po lastnostih pojavnice (*word form* vs. *tags* vs. *lemmas*)
- Sortiranje po KWIC ali levi/desni besedi

Frekvenčni seznami (2/3)



Frekvenčni seznami (3/3)

S CHANGE CRITERIA lahko naredimo frekvenčne sezname znotraj podmnožic v korpusu

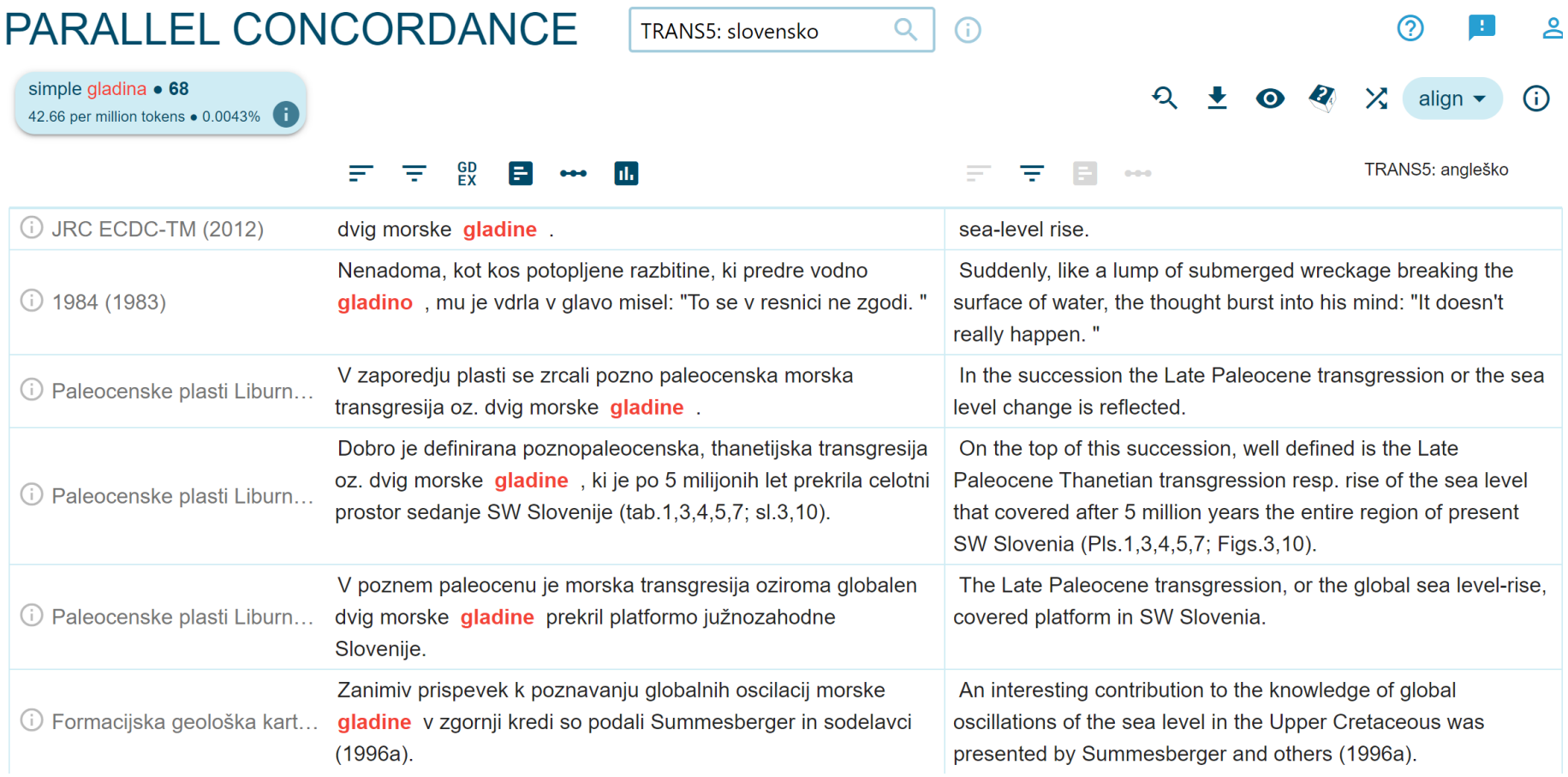


Kolokacije

		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



JRC ECDC-TM (2012)	dvig morske **gladine** .	sea-level rise.
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. "
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.
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).
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.
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. sploh ne stavim na prošnje . jaz g
3	forum • T1 • L3	e živali ! 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	ı.... .... :::)))	**js** jaz	grem pa kr ze dons :P grem pa kar že danes :p še prevoz
5	tweet • T3 • L3	udn zgleda. ıdno zgleda .	**js** jaz	bi tut rabil btw. bi tudi rabil btw . vem pa kako iz pl

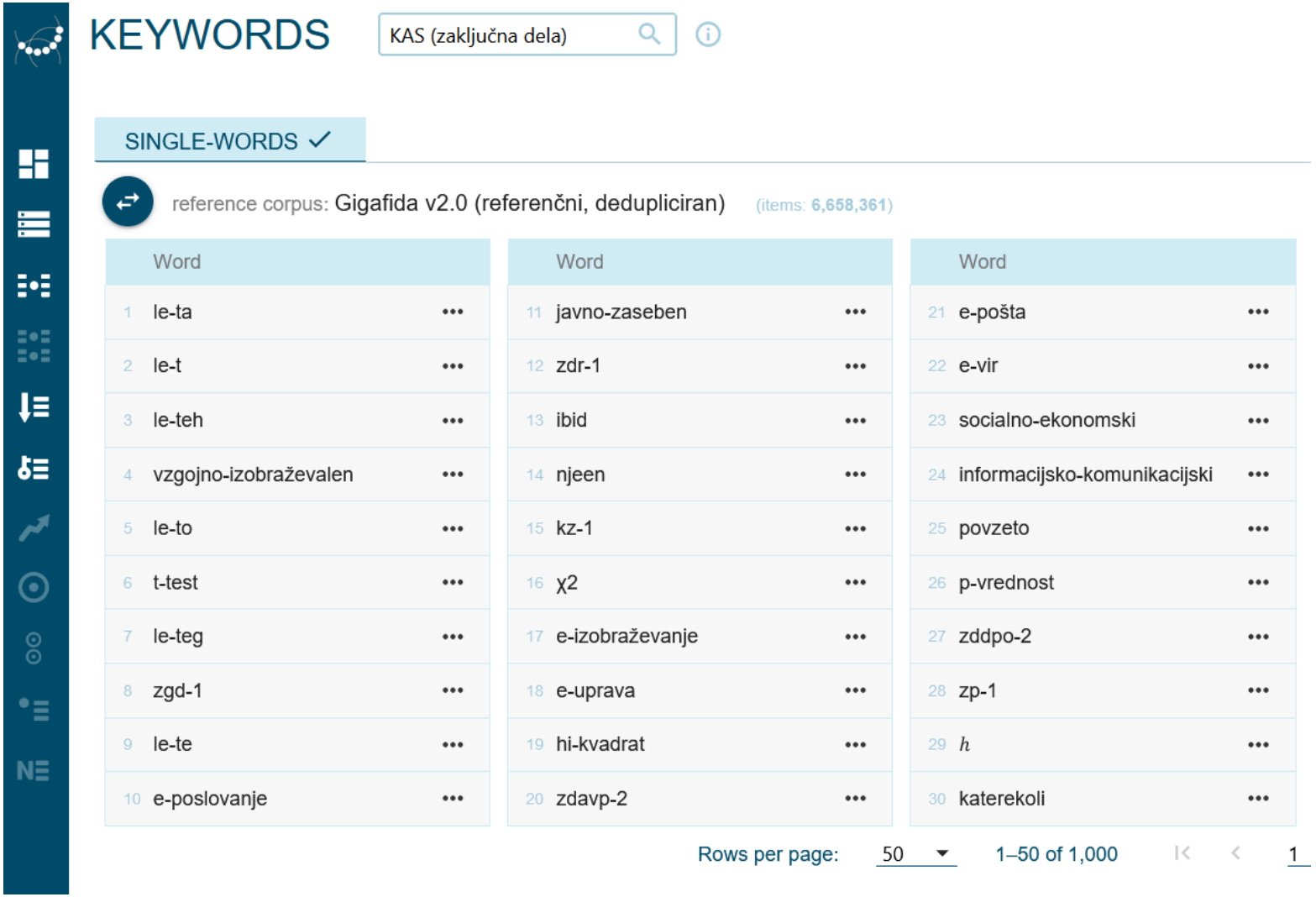
|@iNinaromsek mela saaansooo,
@ininaromsek imela šanso ,
|@nejcjemec to sam pletes?
@nejcjemec to sam pleteš ?
|jaz sploh ne stavim na prošnje.
|vem pa kako iz pl

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 odtreſsite prah od vaſhih nug: Sa riſnizo	jeſt	vam povém, de téh Sodomiterjeu inu Go
		otresite prah od vaših nog : za resnico	jaz	vam povem , da teh sodomiterjev in g
2	Biblija (vzorec...	, na ſodni dan, kakòr takimu Méſtu. Pole,	jeſt	poſhlem vas, kakòr Ouce, v'ſrédo mej V
		, na sodni dan , kakor takemu mesto . pole ,	jaz	pošljem vas , kakor ovce , v sredo med vol
3	Biblija (vzorec...	ganjalitaku beshite v'enu drugu. Riſnizhnu	jeſt	vam povém, vy nebote Israelſka Méſta o
		ganjalitaku bežite v eno drugo . resnično	jaz	vam povem , vi ne_boste izraelska mesta c
4	Biblija (vzorec...	niſhtèr ſkrivniga, kar bi ſe nesvejdilu: Kar	jeſt	vam pravim v'temmi, tu vy pravite na ſvit
		ništer skrivnega , kar bi se ne_zvedelo : kar	jaz	vam pravim v temi , tu vi pravite na sve
5	Biblija (vzorec...	òr veliku Vrabzou. Satu ſlejdni, kateri kuli	mene	ſposná pred Zhloveki, tiga hozhem jeſt ſp
		or veliko vrabcev . zato slednji , kateri koli	mene	spozna pred človeki , tega hočem jaz s
6	Biblija (vzorec...	mene ſposná pred Zhloveki, tiga hozhem	jeſt	ſposnati pred moim Nebeſkim Ozhetom:
		mene spozna pred človeki , tega hočem	jaz	spoznati pred mojim nebeškim očetom :

Ključne besede (1/2)



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	ljubljančanka

Rows per page: 50 1–50 of 1,000 1 / 20

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

Zaključek