Raziskovalna infrastruktura CLARIN.SI

Tomaž Erjavec

Odsek za tehnologije znanja, Institut “Jožef Stefan”
Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU

Predavanje FF
2025-04-14

Pregled predavanja

Uvod

Kje se srečujemo z jezikovnimi podatki

Humanistične vede

Empirično podprte jeziko(slo)vne raziskave:

- temeljijo na realnih besedilih
- za učinkovito uporabo potrebujemo analitična orodja
- jezikoslovci, slovaropisci, zgodovinarji, družboslovci...

Računalništvo

Jezikovne tehnologije:

- obdelava jezika postaja vedno bolj raziskovalno/komercialno zanimivo področje (Google Translate, ChatGPT)
- glavna paradigma: nadzorovano strojno učenje
- taki programi so večinoma jezikovno neodvisni, potrebujejo pa učne (ročno označene) podatke za učenje modela in testne podatke za evalvacijo

Kaj so jezikovni viri

Korpusi

- enovito kodirana in dokumentirana zbirka besedil
- označena (ročno ali strojno)
- referenčni/specializirani; eno/večjezični; pisni/govorni

Leksikalni viri

- besedišča jezika za uporabo v programih
- strojno berljivi slovarji

Modeli jezika

- podatki za nek program, ki mu omogoči obdelavo besedil v nekem jeziku za nek namen
- npr. model CLASSLA za lematizacijo slovenščine; besedne vložitve makedonščine

Ponovna uporaba

Klasični pristop

- za vsako raziskavo izdelati jezikovne vire posebej
- viri nedostopni drugim raziskovalcem

Slabosti

- izdelava jezikovnega vira je lahko zelo draga in dolgotrajna: velika izguba časa in denarja, če se to počne večkrat
- vzdržuje se monopol institucij, ki so vire izdelale
- kasnejši raziskovalci ne morejo preveriti ali poboljšati prvih rezultatov
- viri ne morejo biti uporabljeni pri razvoju produktov

Odprt dostop do rezultatov raziskovalnih projektov

- Brez ovir do publikacij in podatkov
  - prihranek denarja in časa;
  - izogibanje ponavljanju dela;
  - spodbujanje sodelovanja;
  - večja transparentnost znanstvenega procesa;
  - spodbujanje inovacij
- Odprta znanost: močan trend v EU in Sloveniji
- Problemi pri omogočanju odprtega dostopa do jezikovnih virov:
  - avtorske pravice nad besedili
  - varovanje zasebnosti (tudi pravica do pozabe): GDPR
  - pogoji uporabe spletnih portalov (npr. Twitter)

CLARIN(.SI)

Raziskovalne infrastrukture

Kaj je RI?

Naprave, podatki in storitve, ki jih znanstvena skupnost uporablja pri raziskovanju na svojem področju.

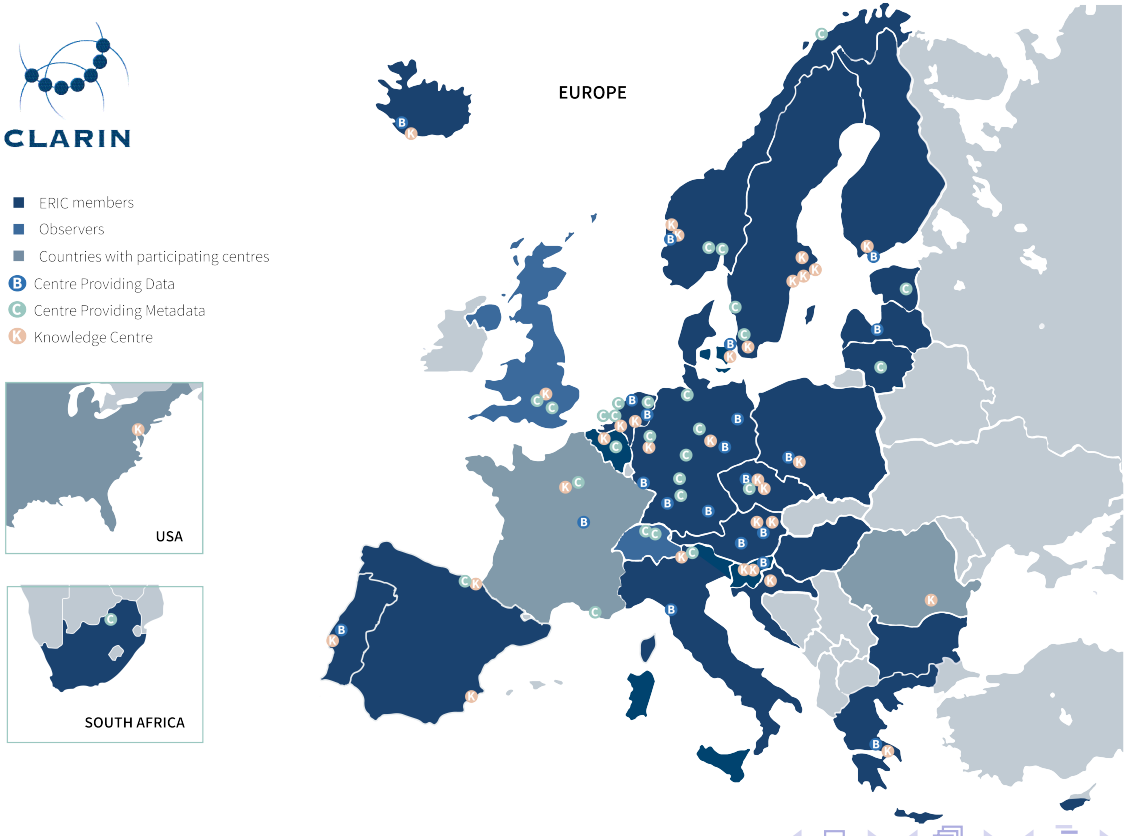
- ESFRI: European Strategy Forum on Research Infrastructures
- Razvojni načrti: 2006 (35 RI), ..., 2018 (55), 2021 (66)
- 22 RI je organiziranih kot ERIC
  (European RI Consortium = evropska pravna oseba)
- Slovenija sodeluje v 22 RI, 2 s področja humanistike:
- DARIAH ERIC / DARIAH-SI = Digital Research Infrastructure for the Arts and Humanities / Digitalna raziskovalna infrastruktura za umetnost in humanistiko: INZ + ZRC SAZU
- **CLARIN ERIC** / **CLARIN.SI** = Common Language Resources and Technology Infrastructure / Infrastruktura za jezikovne vire in tehnologije

CLARIN: Common Language Resources and Technology Infrastructure

- Vizija: digitalni jezikovni viri in orodja za vse (evropske) jezike so dostopni prek enotne prijave za raziskovalce v humanistiki in družboslovju
- Namenjena je dolgotrajnemu in obsežnemu hranjenju ter dostopu do jezikovnih virov in tehnologij
- Prispevek k ohranjanju in podpiranju večjezične evropske kulturne dediščine
- Paradigma sodelovanja pri razvoju virov in orodij, zagotavljanje večkratne uporabnosti in prilagajanja individualnim potrebam

CLARIN ERIC

- Sedež na Nizozemskem
- Direktorica: dr. Daja Fišer, INZ
- Trenutno 24 držav članic + 2 opazovalki
- Podporno osebje, odbori za upravljanje, delovne skupine
- Večina dela se odvija v okviru nacionalnih konzorcijev



Kaj CLARIN ERIC ponuja slovenskim raziskovalcem?

- S slovensko EduGain prijavo dostop do vseh virov in storitev centrov CLARIN držav članic
- Spletne storitve, npr. virtualni jezikovni observatorij
- Podpora ciljnim projektom, npr. razvoju učnih vsebin, izvedbi delavnic za uporabnike, snovanju evropskih projektnih prijav
- Infrastruktura znanja:

CLARIN.SI

- Začetek dela v 2015
- Institut "Jožef Stefan":
  - Odsek za tehnologije znanja (E8)
  - Laboratorij za umetno inteligenco (E3)
  - Center za mrežno infrastrukturo (CMI)
- Organiziran kot konzorcij 12 partnerjev:
  - 4 univerze: Ljubljana, Maribor, Nova Gorica, Primorska
  - 4 raziskovalni inštituti: ZRC SAZU, IJS, INZ, ZRS Koper
  - 1 knjižnica: NUK
  - 1 društvo: Slovensko društvo za jezikovne tehnologije
  - 2 podjetji: Amebis, Alpineon

Storitve

Delovanje CLARIN.SI

Trije stebri

1. Repozitorij jezikovnih virov in orodij
2. Spletne storitve (glavna: **konkordančniki**)
3. Podpora digitalni humanistiki in jezikovnim tehnologijam (prenos znanja, dogodki, projekti)

Repozitorij jezikovnih virov in orodij

- Trenutno najpomembnejša storitev CLARIN.SI
- Arhiv $>$ 650 jezikovnih virov in orodij, od tega 350 (tudi) za slovenski jezik: korpusi, slovarji, besedišča, modeli, programi
- Samoarhiviranje + uredniški pregled
- Repozitorij certificiran s strani CLARIN in Core Trust Seal
- Dolgotrajno hranjenje, avtentikacija in avtorizacija, stalni identifikatorji, eksplicitni pogoji uporabe in licence
- Pomemben doprinos k odprti znanosti

Konkordančniki CLARIN.SI

- Orodja za (spletno) analizo korpusov
- Besede v kontekstu, frekvenčni seznami, ključne besede...
- Podpirajo delo z zelo velikimi korpusi (več milijard besed)
- Korpusi so lahko bogato označeni
- Zmogljiv poizvedovalni jezik
- Raznovrstni izpisi in analize
- Ponujajo cca. 100 korpusov v 33 jezikih z 20 milijard besed

Centri znanja

- CLARIN certificira centre znanja (K-centres), ki nudijo strokovno podporo za področna, povezana z delovanjem RI
- CLARIN.SI sodeluje v več centrih znanja:
  - CLASSLA: obdelava južnoslovanskih jezikov
  - ELEXIS: digitalna leksikografija
  - LLMs4SSH veliki jezikovni modeli v humanistiki in družboslovju
  - CKCMC: računalniško posredovana komunikacija in korpusi družbenih omrežij
- CLARIN.SI ustanovitelj CLASSLA:
  - Tehnična podpora pri izdelavi in uporabi virov in orodij
  - Pogosto zastavljena vprašanja za slv., hrv., srb., bolg. in mak.
  - Izdelava korpusov in orodij
  - CLASSLA Express serija delavnic o uporabi spletnih korpusov:
    2024 (Zagreb, Reka, Beograd, Skopje, Sofija, Ljubljana)
    2025 (Celovec, Zagreb, Gradec, Reka, Bled)

Strokovna podpora in diseminacija

- Od 2018 letna finančna podpora projektom, letno izbranih na odprtem razpisu za člane konzorcija (36 uspešno zaključenih)
- Organizacija in podpora dogodkom, predvsem mednarodni konf. Jezikovne tehnologije in digitalna humanistika (na dve leti od 1998)
- Obveščanje in promocija (predstavitve na konferencah, študentom, izvedba tečajev)

Vpetost v projekte in RI

- MIZŠ (2018–2021): nadgradnja strojne opreme
- EU ELEXIS (2018–2022): repozitorijska zbirka metapodatkov 143 digitalnih slovarjev
- MK RSDO (2020–2023): pregled in arhiviranje jezikovnih virov projekta
- RI CESSDA/ADP RDA Node Slovenia (2019–2020): pregled in analiza slovenskih repozitorijev raziskovalnih podatkov
- RI DARIAH-SI/INZ: sodelovanje na področju standardizacije zapisa in izdelave korpusov parlamentarnih podatkov
- CLARIN ERIC:
  - 2 manjša projekta (2016, 2019) + mednarodni delavnici
  - ključna vloga v “CLARIN Flagship” projektih: ParlaMint I (2020–2021), ParlaMint II (2022–2023): izdelava korpusov parlamentarnih razprav
  - več nagrad in priznanj slovenskim znanstvenikom za delo, povezano s CLARIN

Konkordančniki

Na splošno o konkordančnikih

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

vrka med vejicami in kupom dračja kot	**miška**	. Je žužkojed. Med drozgi lahko ž
ziran v slovenščino. Pripravite se mala	**miška**	bo doživela veliko pustolovščino.
anuarja ob 18.30 in 10. februarja v OŠ	**Miška**	Kranjca v Ljubljani ob 17.30. Zara
: iste stranke neuradno omenja Andrej	**Miška**	, ki si je s sedanjim županom v sp
olju kot prilogo nshimi. V vedrih ocvrte	**miške**	so najljubša lokalna sladica. D4) Z
tnimi raziskavami na glodavcih, kot so	**miške**	in podgane, so tiste z ribami cebri
i pri nakupu tablice dodatno dobili tudi	**miško**	. Posebno pri cenejših tablicah zn
ıo ali bi radi počeli tudi legendarni Miki	**Miška**	in njegova izvoljenka! Nikakor var

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

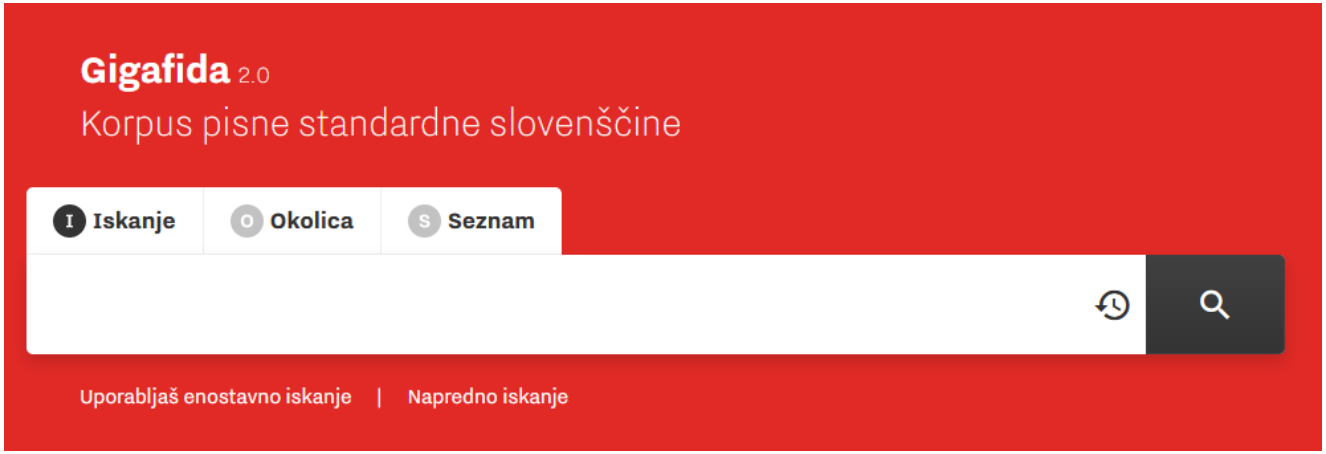
“English Corpora” (https://www.english-corpora.org/)

- 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, finščina)

Konkordančniki v Sloveniji

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



- Drugi konkordančniki CJVT: Šolar, Lektor, GOS
- Drugi namenski konkordančniki: *EvroTerm*

Konkordančniki CLARIN.SI

- Besede v kontekstu, frekvenčni seznami, ključne besede, kolokacije, ...
- Podpirajo delo z zelo velikimi korpusi (več milijard besed)
- Korpusi so lahko bogato označeni:
  - strukture: besedilo, odstavek, termin . . .
  - metapodatki: leto izdaje, vrsta besedila, spol avtorja . . .
  - atributi pojavnic: **oblikoskladenjska oznaka, lema**, normalizirana oblika . . .
- Bogat poizvedovalni jezik
- Raznovrstni izpisi in analize
- RESTful vmesnik, tj. poizvedbe možne prek URLjev (rezultat lahko tudi v JSON ali XML)
- Ponujajo cca. 100 korpusov v 33 jezikih z 20 milijard besed

O nekaj pomembnejših korpusih (1/3)

Referenčni korpus **GigaFida**

- Pisna standardna slovenščina
- 1.1 milijarda besed
- Besedila (predvsem tisk) 1990–2018

Korpusi spletne komunikacije JANES

- Tviti, forumi, blogi, komentarji na novice
- 250 milijonov besed
- Pomembni za raziskavo spletnega jezika
- Posebne oznake (normalizacija)

O nekaj pomembnejših korpusih (2/3)

Korpusi starejše slovenščine IMP

- Razpon: 16. stoletje – 1918
- 15 milijonov besed
- Verske knjige, leposlovje, učbeniki, Kmetijske in rokodelske novice, itd.
- Posebne oznake (normalizacija)

Korpus akademske slovenščine KAS

- Diplomske naloge, magistrska dela, doktorske disertacije
- 1,3 milijarde besed
- Pomemben vir za preučevanje akademske slovenščine
- Označeni termini
- Novejši **Korpus odprte znanosti OSS**: večji (2,4 milijarde besed) z več vrst besedil, a bolj umazan

O nekaj pomembnejših korpusih (3/3)

Slovenske parlamentarne razprave siParl

- Razpon: 1990 – 2022
- 230 milijonov besed
- Bogati metapodatki o govorcih
- Večjezični korpusi **ParlaMint**: 29 evropskih parlamentov, tudi avtomatsko prevedeni v angleščino (1,5 milijarde besed)

Združeni korpus MetaFida

- Sestavljen iz 33 korpusov
- Največji korpus slovenščine: 3,5 milijarde besed
- Vsebuje samo oznake, ki so skupne vsem vsebovanim korpusom

Vsi konkordančniki CLARIN.SI

noSketch Engine (Bonito)

- Stara verzija nekomercialnega noSketch Engine
- Lexical Computing ga ne vzdržuje več, ampak:
- Še vedno popularen na CLARIN.SI

noSketch Engine (Crystal) $\Leftarrow$ danes podrobneje o temle

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

NoSketch Engine Crystal s prijavo

- Potrebna prijava (samoregistracija)
- Zato pa omogoča izdelavo podkorpusov, prilagoditve pogleda, zgodovino poizvedb

KonText

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

Uporaba noSketch Engine

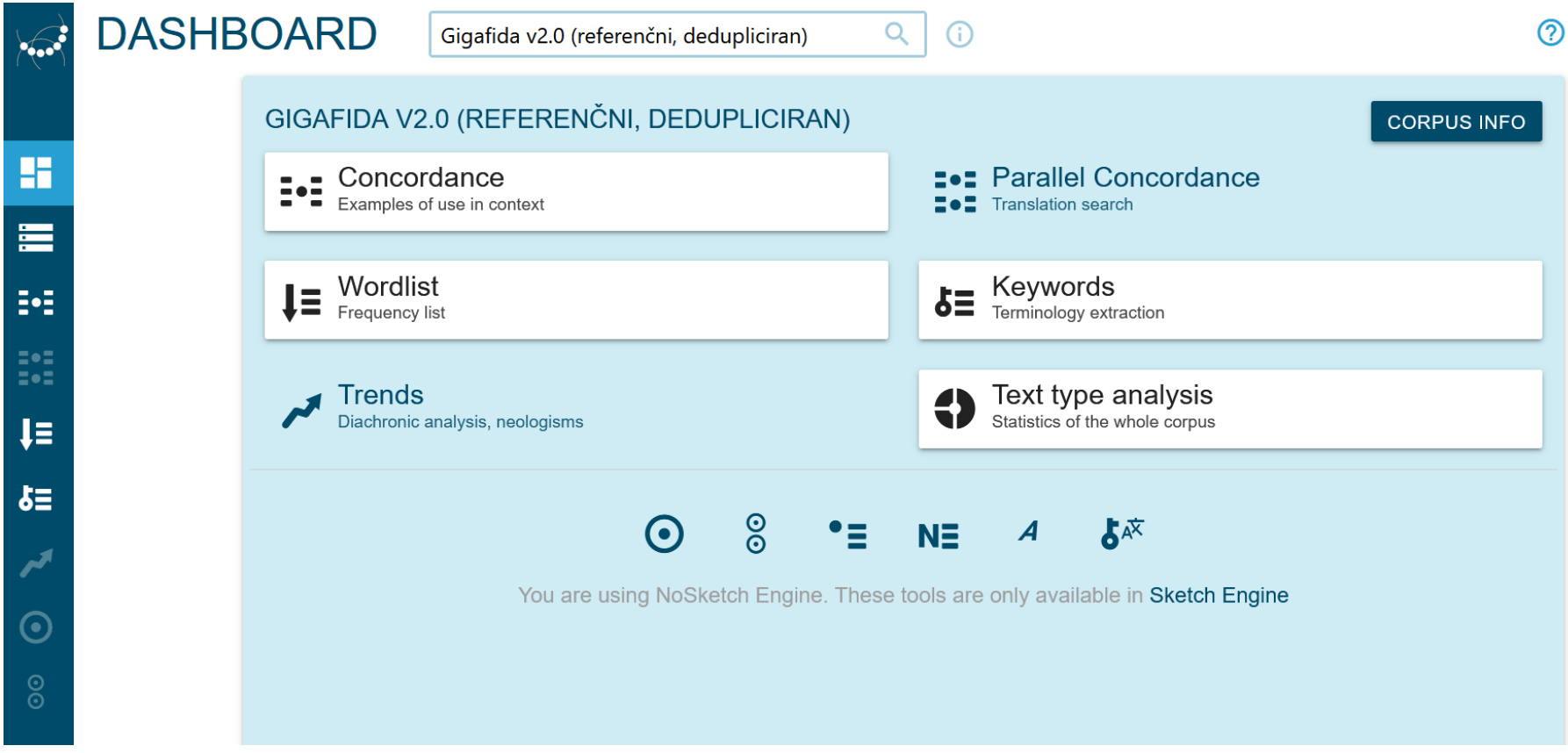
noSketch Engine (Crystal)

https://www.clarin.si/ske

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

noSke nadzorna plošča



- 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

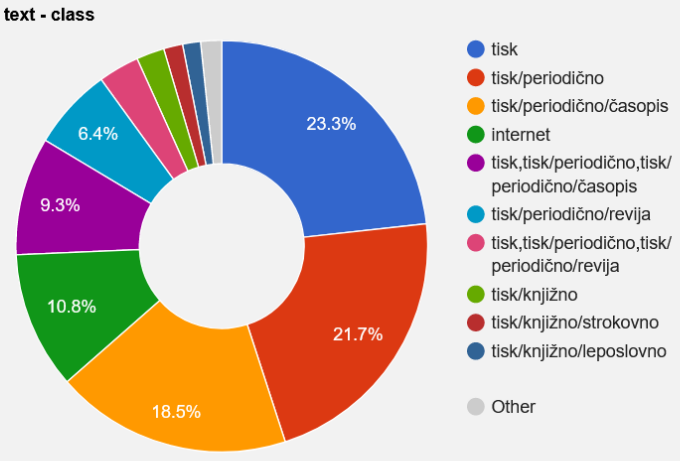
- text - author
- **text - class**
- text - date
- text - id
- text - publisher
- text - source
- text - title

Show: Token coverage

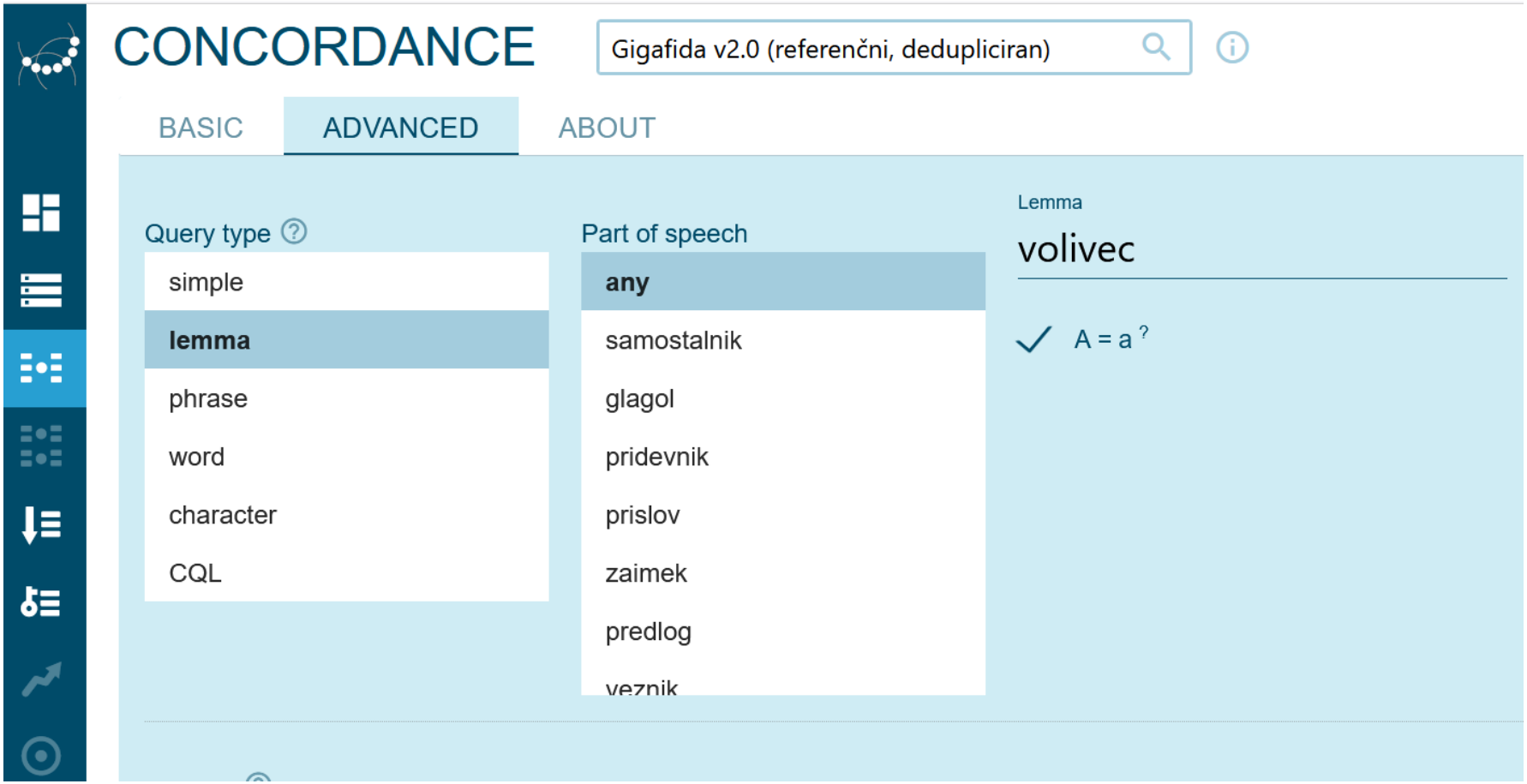
Subcorpus: none (the whole corpus)

Filter results

Items: **15**, Total frequency: **6,873,401,581**

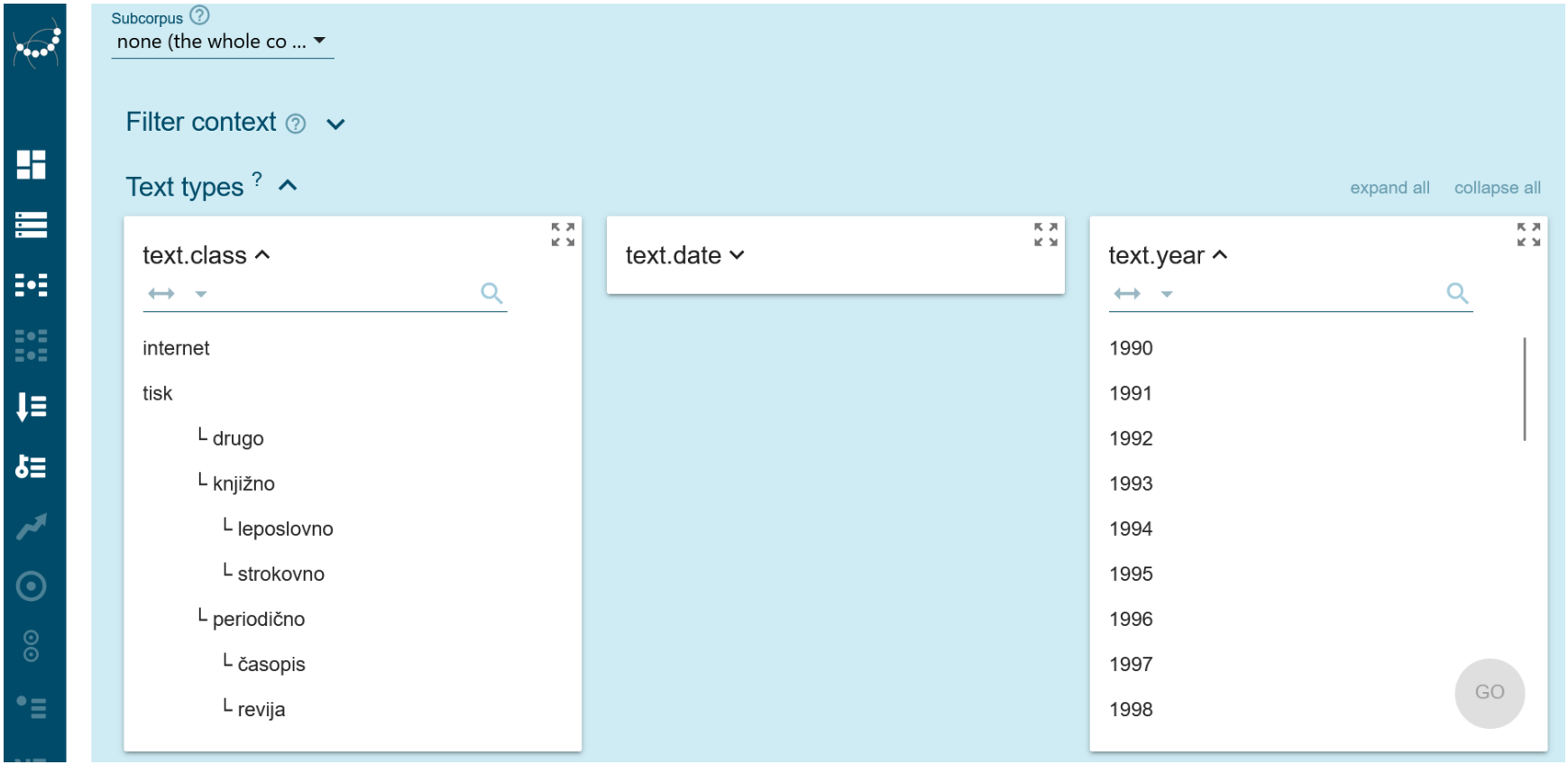


Konkordance (1/3)



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

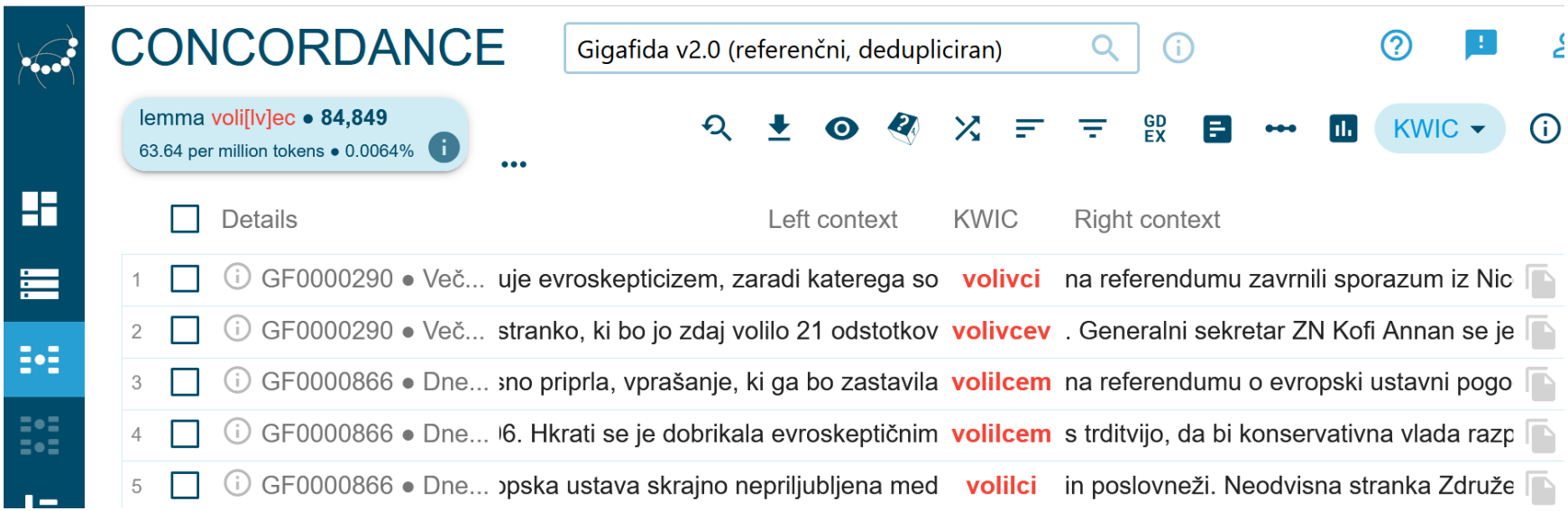
Konkordance (2/3)



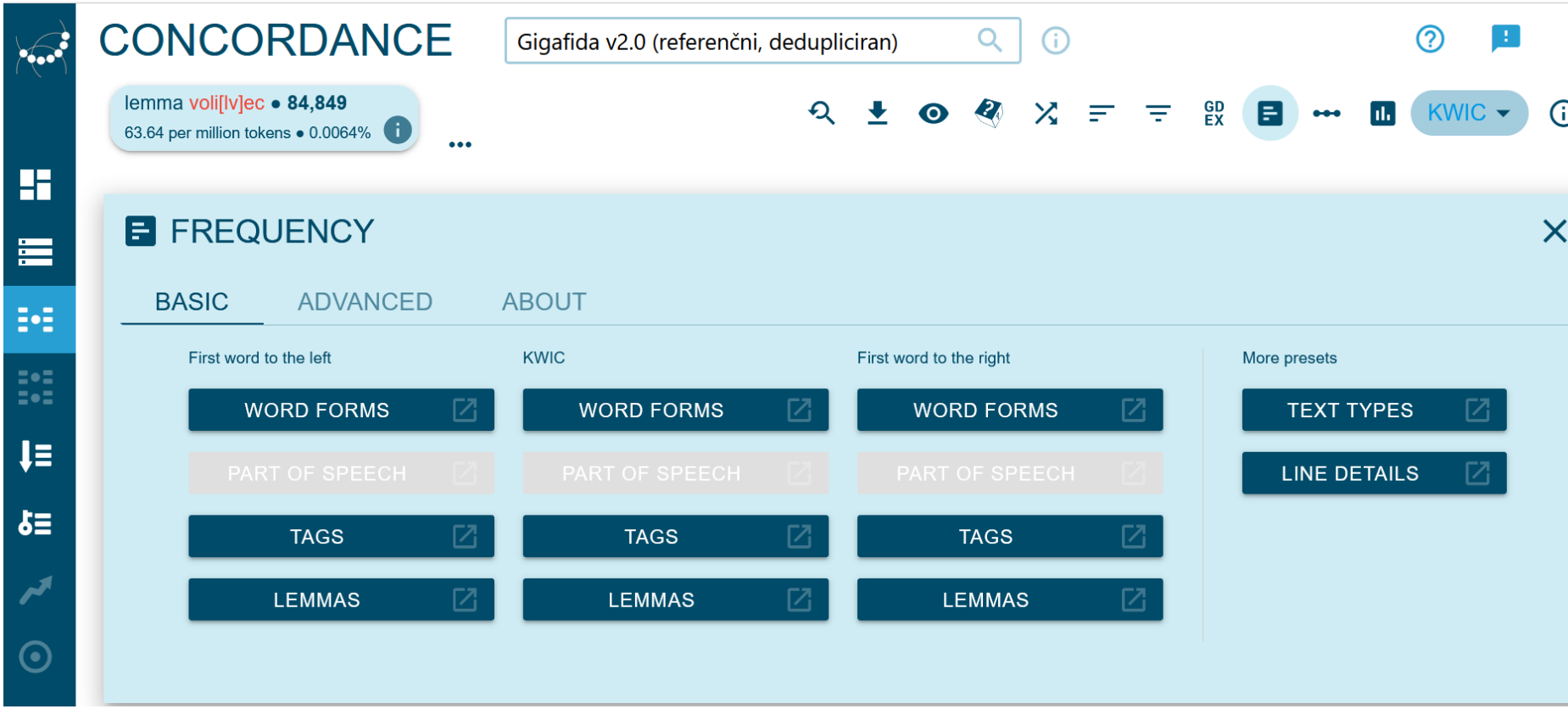
- Omejevanje iskanja na dele korpusa
- V primeru GigaFide: leto, besedilna vrsta, avtor ipd.

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)

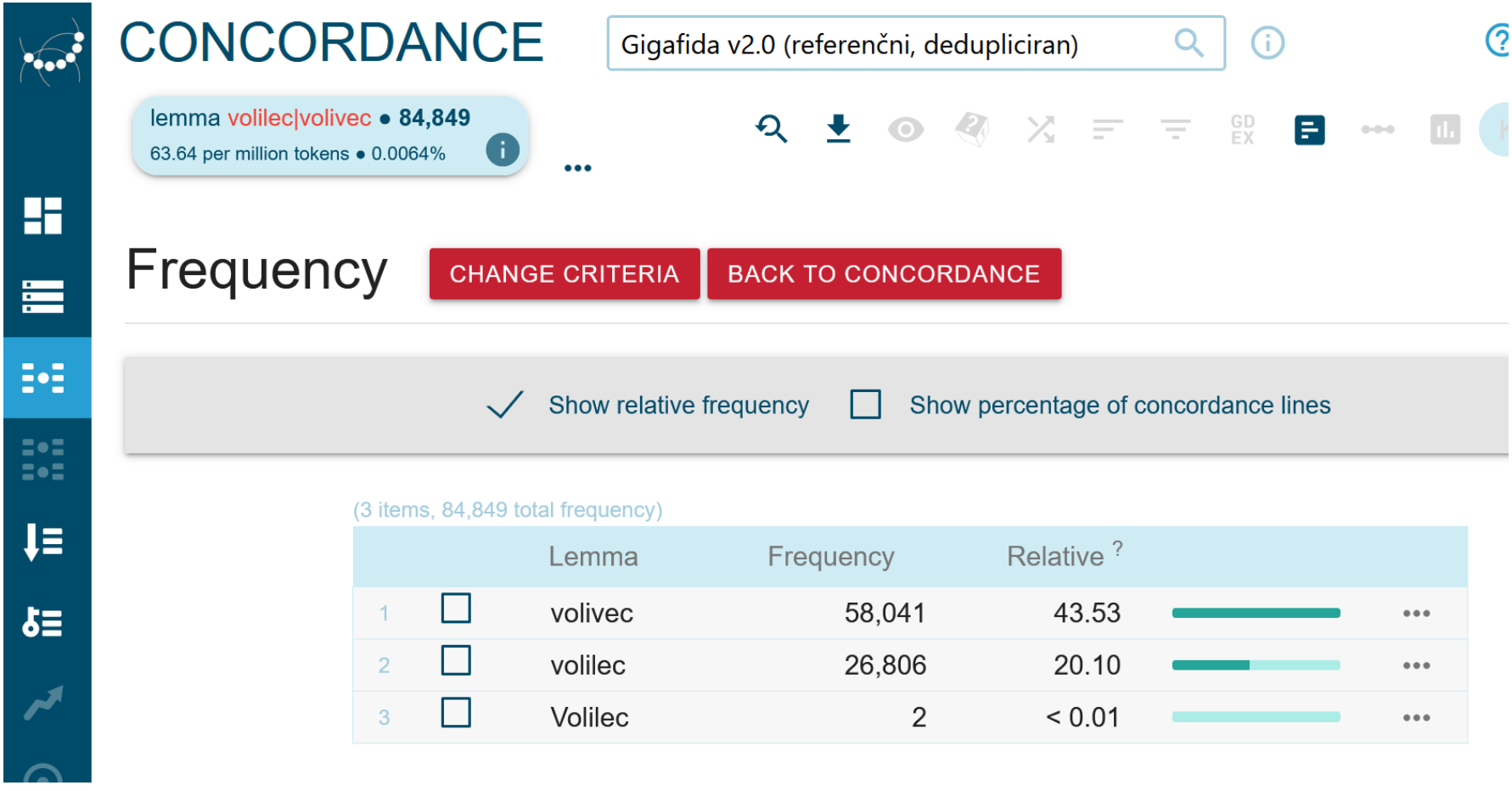


Frekvenčni seznami (1/3)



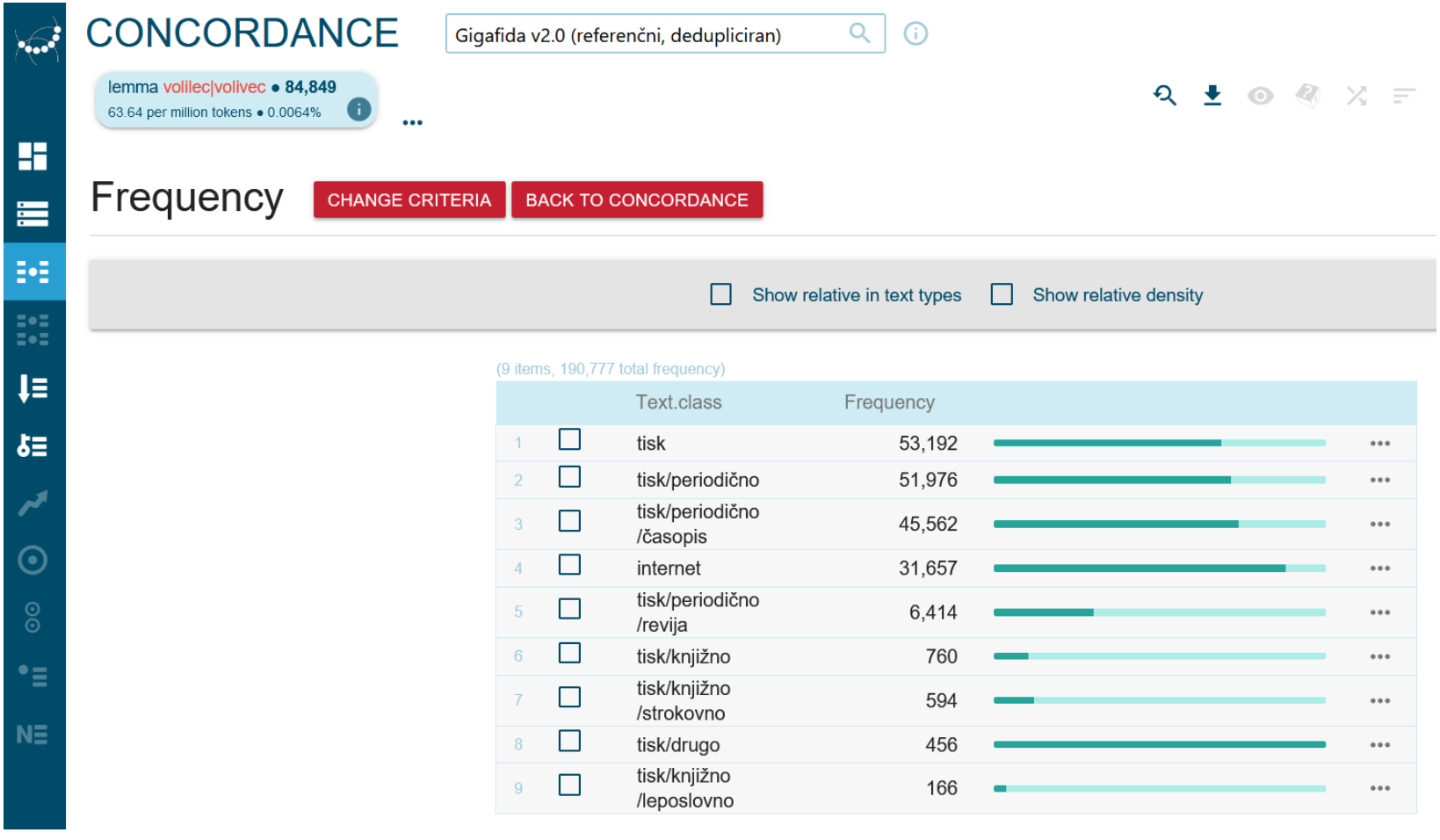
- Sortiranje po lastnostih pojavnice (*word form* vs. *tags* vs. *lemmas*)
- Sortiranje po KWIC ali levi/desni besedi

Frekvenčni seznami (2/3)



Frekvenčni seznami (3/3)

S CHANGE CRITERIA lahko naredimo frekvenčne sezname znotraj podmnožic v korpusu

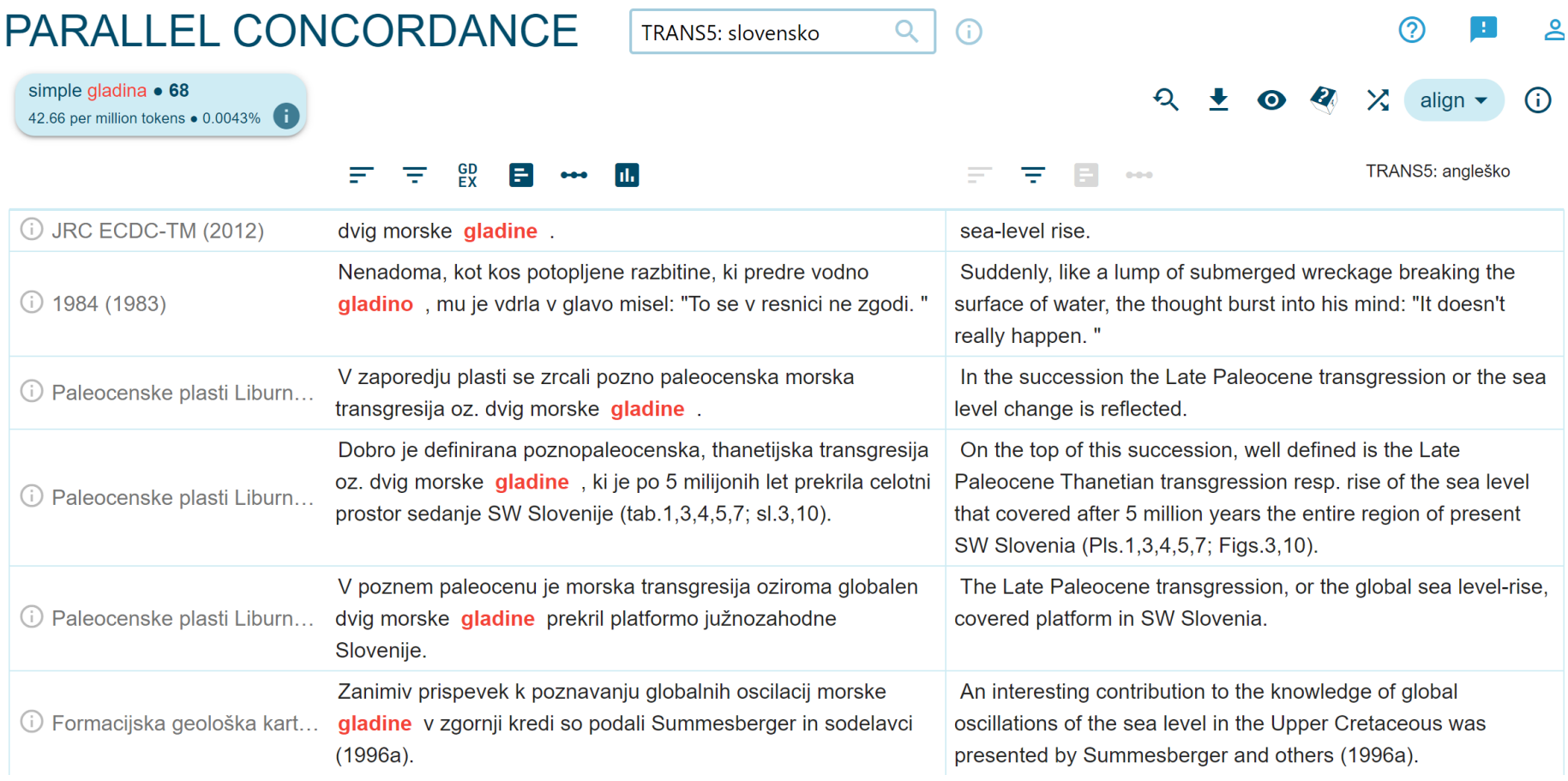


Kolokacije

		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



PARALLEL CONCORDANCE

TRANS5: slovensko

simple gladina • 68
42.66 per million tokens • 0.0043%

align

TRANS5: angleško

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*

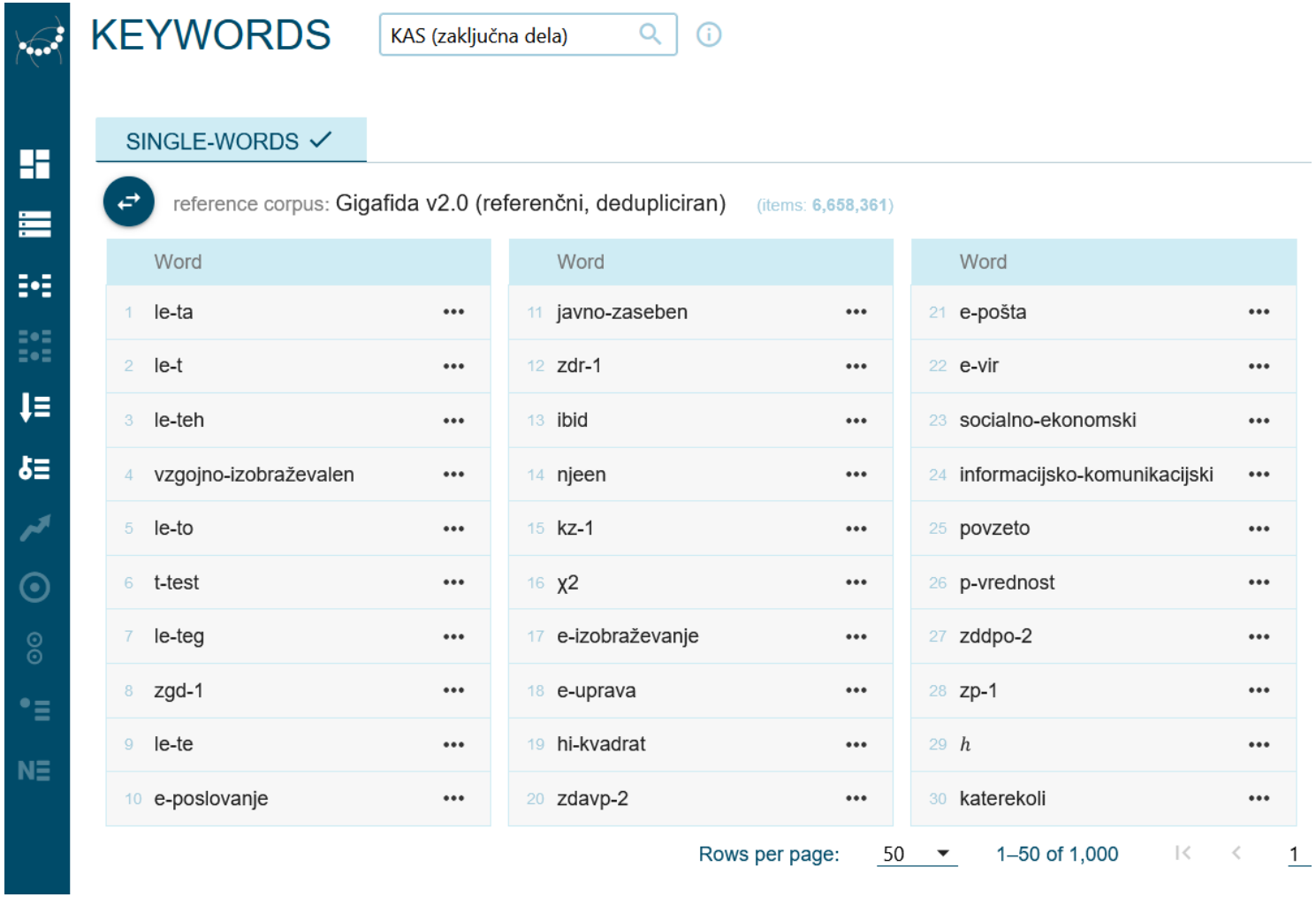
	Details	Left context	KWIC	Right context
1	tweet • T3 • L3	|@spirulinka9 @tretjeoko pri ex sem si že @spirulinka9 @tretjeoko pri ex sem si že	**jaz** jaz	tud marsikaj lahko predstavljala..| tudi marsikaj lahko predstavljala ..
2	forum • T1 • L3	pa lova na kite in ostale nedolžne živali !| pa lova na kite in ostale nedolžne živali !	**Jaz** jaz	sploh ne stavim na prošnje.|Jaz g sploh ne stavim na prošnje . jaz g
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	ı....|:::)))|@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?| udno zgleda . @nejcjemec to sam pleteš ?	**js** jaz	bi tut rabil btw.|vem pa kako iz ple bi tudi rabil btw . vem pa kako iz ple

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

Ključne besede (1/2)



Ključne besede v Korpusu akademske slovenščine glede na GigaFido

Ključne besede (2/2)

KEYWORDS

Gigafida v2.0 (referenčni, dedupliciran)

SINGLE-WORDS ✓

reference corpus: KAS (zaključna dela) (items: 3,158,022)

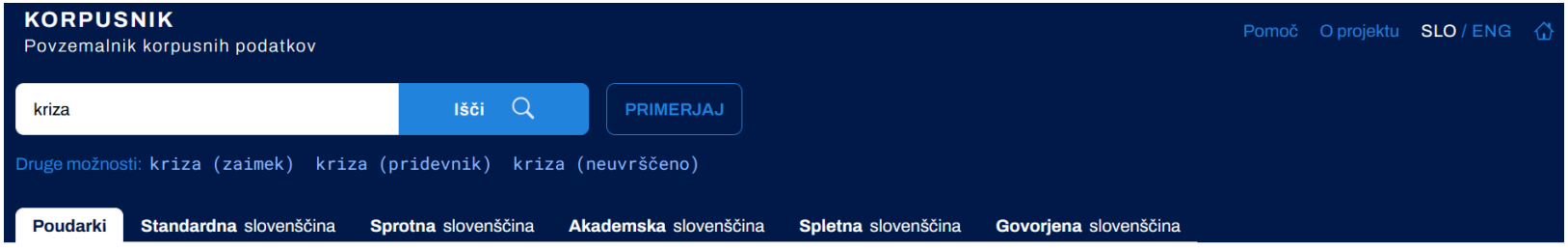
	Word		Word		Word		Word		Word
1	ugnati	11	kolajna	21	đoković	31	četrtkov	41	priigrati
2	oblačno	12	dirkač	22	letos	32	torkov	42	pokal
3	tisočak	13	polfinale	23	četrtfinale	33	remi	43	selektor
4	včeraj	14	sinoči	24	podprvak	34	koprčan	44	hina
5	derbi	15	lani	25	prvakinja	35	messi	45	hokejist
6	prvoligaš	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

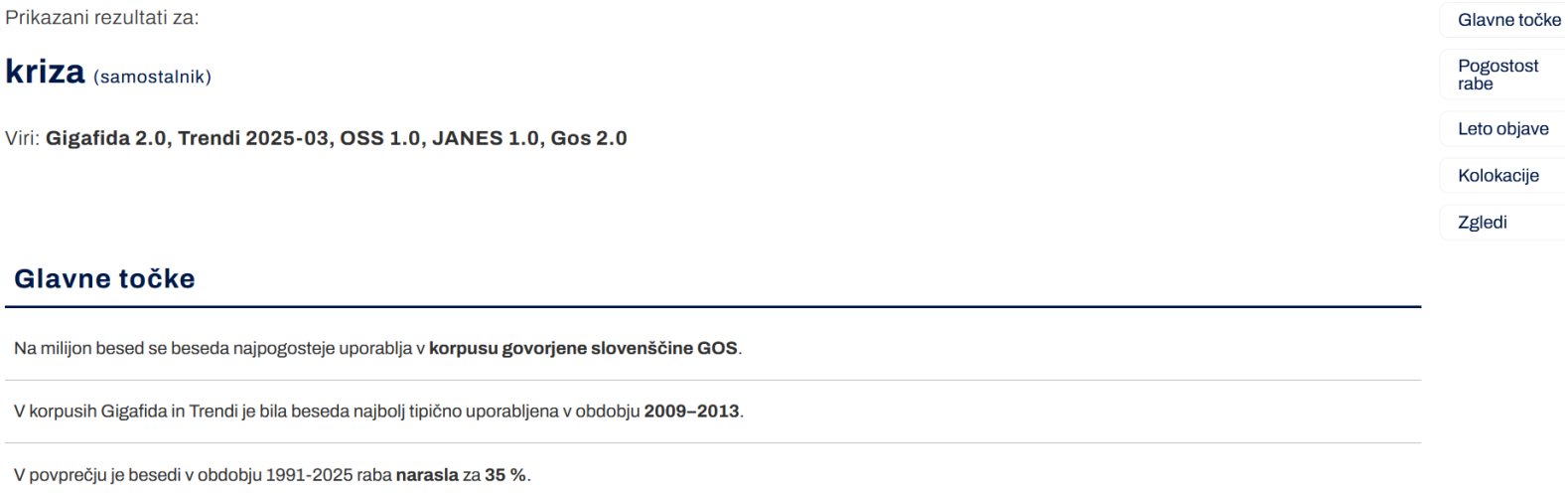
Rows per page: 50 1–50 of 1,000 1 / 20

Ključne besede v GigaFidi glede na Korpus akademske slovenščine

Korpusnik

- Spletna storitev CJVT: povzemalnik korpusnih podatkov
- Konkordančnike približa širši množici uporabnikov
- Poizvedbo pošlje na več korpusov in povzame rezultate





Zaključek

Zaključki

Na kratko

CLARIN.SI nudi možnost trajnega arhiviranja jezikovnih virov, odprt in brezplačen dostop do jezikovnih virov, orodij in storitev za (slovenske) raziskovalce in (kjer le mogoče) podjetja ter podporo pri ustvarjanju, arhiviranju in uporabi jezikovnih virov in orodij.

Nadaljnje informacije

- https://www.clarin.eu/
- https://www.clarin.si/
- Pričujoče prosojnice: https://www.clarin.si/info/dogodki/
- ERJAVEC, Tomaž, DOBROVOLJC, Kaja, FIŠER, Darja, JAVORŠEK, Jan Jona, KREK, Simon, KUZMAN, Taja, LASKOWSKI, Cyprian Adam, LJUBEŠIĆ, Nikola, MEDEN, Katja. Raziskovalna infrastruktura CLARIN.SI. V: Jezikovne tehnologije in digitalna humanistika: zbornik konference. 2022. https://doi.org/10.5281/zenodo.14165471